AMERICAN ACADEMY OF OPHTHALMOLOGY®

Improving the Identification of Diabetic Retinopathy and Related Conditions in the Electronic Health Record Using Natural Language Processing Methods

Keith Harrigian, MS,[1] Diep Tran, MSc,[2] Tina Tang, MD,[2] Anthony Gonzales, OD,[2] Paul Nagy, PhD,[3] Hadi Kharrazi, MD, PhD,[4] Mark Dredze, PhD,[1] Cindy X. Cai, MD, MS[2,3]

Purpose: To compare the performance of 3 phenotyping methods in identifying diabetic retinopathy (DR) and related clinical conditions.

Design: Three phenotyping methods were used to identify clinical conditions including unspecified DR, nonproliferative DR (NPDR) (mild, moderate, severe), consolidated NPDR (unspecified DR or any NPDR), proliferative DR, diabetic macular edema (DME), vitreous hemorrhage, retinal detachment (RD) (tractional RD or combined tractional and rhegmatogenous RD), and neovascular glaucoma (NVG). The first method used only International Classification of Diseases, 10th Revision (ICD-10) diagnosis codes (*ICD-10 Lookup System*). The next 2 methods used a Bidirectional Encoder Representations from Transformers with a dense Multilayer Perceptron output layer natural language processing (NLP) framework. The NLP framework was applied either to free-text of provider notes (*Text-Only NLP System*) or both free-text and ICD-10 diagnosis codes (*Text-and-International Classification of Diseases* [*ICD*] *NLP System*).

Subjects: Adults ≥18 years with diabetes mellitus seen at the Wilmer Eye Institute.

Methods: We compared the performance of the 3 phenotyping methods in identifying the DR related conditions with gold standard chart review. We also compared the estimated disease prevalence using each method.

Main Outcome Measures: Performance of each method was reported as the macro F1 score. The agreement between the methods was calculated using the kappa statistic. Prevalence estimates were also calculated for each method.

Results: A total of 91 097 patients and 692 486 office visits were included in the study. Compared with the gold standard, the *Text-and-ICD NLP System* had the highest F1 score for most clinical conditions (range 0.39–0.64). The agreement between the *ICD-10 Lookup System* and *Text-Only NLP System* varied (kappa of 0.21–0.81). The prevalence of DR and related conditions ranged from 1.1% for NVG to 17.9% for DME (using the *Text-and-ICD NLP System*).

Conclusions: The prevalence of DR and related conditions varied significantly depending on the methodology of identifying cases. The best performing phenotyping method was the *Text-and-ICD NLP System* that used information in both diagnosis codes as well as free-text notes.

Financial Disclosures: Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100578 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Diabetes mellitus is a leading cause of vision loss among working age adults in the United States.[1] As the prevalence of diabetes rises, diabetic eye diseases including retinopathy, macular edema (ME), and related conditions (e.g., vitreous hemorrhage [VH], tractional retinal detachments [TRDs], and neovascular glaucoma [NVG]) are also expected to increase.[2] Despite the rising prevalence, identifying such patients in routinely collected observational health data, for example, the electronic health record (EHR), is challenging. Accurately identifying such patients is important for establishing disease prevalence for epidemiologic surveillance or public health planning, for example, to ensure there are sufficient health care providers to treat the conditions in certain areas.[3] The complexity and richness of the data available in the EHR also enables hypothesis driven precision medicine research that accounts for the uniqueness of each individual.[4]

https://doi.org/10.1016/j.xops.2024.100578
ISSN 2666-9145/24

Accurate case identification of patients with diabetic retinopathy (DR) and related conditions is essential to studies using routinely collected observational health data. Many studies rely on structured data including International Classification of Diseases (ICD) diagnosis codes for case identification.[5,6] Although investigators have found acceptable concordance between billing and provider free-text documentation for many ocular conditions, there are limitations to solely relying on ICD diagnosis codes for case identification.[7,8] The accuracy of ICD diagnosis codes can vary by condition and is influenced by local billing practices and documentation workflows.[9,10] For example, evidence suggests that more severe clinical diseases that necessitate changes in clinical management, such as proliferative DR (PDR), tend to be captured more completely than less severe diseases like nonproliferative DR (NPDR).[9] There are also missing data in diagnosis codes selected that do not specify laterality (i.e., which eye is affected) or severity (e.g., mild or severe NPDR).[9] Nonspecific diagnosis codes are seen more commonly for older patients, those with better vision, patients with lower utilization of eye care, and also depends on the specialty of the eye care provider.[11]

Sole reliance on ICD diagnosis codes could be misrepresenting the identification of DR and related conditions. Electronic health record data represent a unique opportunity to compare diagnosis codes with free-text provider notes of clinical conditions. To address the identification gap of DR and related conditions using EHR data, this study aimed to (1) develop and share a natural language processing (NLP) framework that can identify DR and related conditions in ophthalmic provider free-text notes, (2) compare the identification of these clinical conditions between the provider free-text notes and structured ICD diagnosis codes, and (3) report the performance of a combined NLP and ICD method for optimal sensitivity and specificity in identifying DR and related conditions.

Methods

Adult patients ≥18 years with diabetes mellitus seen at a tertiary care academic referral center (Wilmer Eye Institute) were included in the study.[12] All ophthalmology office visits containing free-text notes from January 1, 2013 to April 1, 2022 were identified. Ten DR related clinical conditions of interest were included: unspecified DR, NPDR (mild, moderate, severe), consolidated NPDR (unspecified DR or any NPDR), PDR, diabetic macular edema (DME), VH, retinal detachment (RD) (TRD or combined tractional and rhegmatogenous RD), and NVG. The extraction of these clinical concepts and associated attributes (e.g., laterality) were compared across 3 phenotyping methods: International Classification of Diseases, 10th Revision (ICD-10) diagnosis codes (*ICD-10 Lookup System*), NLP framework applied to provider free-text notes (*Text-Only NLP System*), and NLP framework applied to free-text notes and ICD-10 codes (*Text-and-ICD NLP System*). The performance of these systems were evaluated against a gold standard. The study was approved by the Johns Hopkins Institutional Review Board which waived the requirement for patient consent as this was secondary research. The study also adhered to the tenets of the Declaration of Helsinki.

Gold Standard

Free-text notes (i.e., progress notes and problem list documentation) and encounter-level ICD-10 codes were extracted from EHR's data warehouse and merged using a standardized template into a formatted note (Fig S1, available at www.ophthalmologyscience.org). Formatted notes from a random selection of 736 office visits from 348 patients were annotated to establish the gold standard labels for training and validating the NLP framework, and validating the *ICD-10 Lookup System*. A set of high-recall regular expressions was used to identify candidate text spans indicating the clinical concepts in the formatted notes (e.g., "retinal tear" to identify RD). The regular expressions were curated based on a combination of domain knowledge and a qualitative review of 50 randomly sampled notes from the validation split. Details of the precise regular expressed used can be found in the GitHub link.[13] The text spans were annotated by 2 graders (a postgraduate year-4 ophthalmology resident [T.T.] and a licensed optometrist [A.G.]) for correctness (i.e., whether the text span indicates the intended clinical concept), and to assign attribute labels (laterality, status, and severity/type) (Table S1, available at www.ophthalmologyscience.org). A total of 3042 text spans were annotated, with a total of 8973 attribute labels. Disagreements were resolved through discussion overseen by a board-certified ophthalmologist and clinical informaticist (C.X.C.). A detailed description of the annotation process can be found in the appendix of Harrigian et al.[14]

Span-Level to Encounter-Level Condition Resolution

The gold standard annotation and NLP framework prediction were completed using the span-level text, thus incongruent labels were observed on the encounter-level (e.g., simultaneous identification of mild NPDR and PDR in the right eye in a given office visit). To address this issue, we applied post hoc resolution logic to the gold standard, *ICD-10 Lookup System*, *Text-Only NLP System*, and *Text-and-ICD NLP System* (Fig S2, available at www.ophthalmologyscience.org). The resolution process assigned the highest identified severity of each condition. For DR, severity from least to most severe was as follows: unspecified DR, mild NPDR, moderate NPDR, severe NPDR, and PDR. The assignment was done for each eye (designed the "per-eye" resolution), and for each person (designed the "per-person" resolution).

Phenotyping Methods

Three phenotyping methods (i.e., *ICD-10 Lookup System*, *Text-Only NLP System*, and *Text-and-ICD NLP System*) of extracting DR related clinical conditions and associated attributes were developed and compared. A comparison of the 3 methods on a synthetic formatted note from a single office encounter is shown in the supplement (Fig S3, available at www.ophthalmologyscience.org).

ICD-10 Lookup System

We extracted the ICD-10 encounter diagnoses associated with the ophthalmology office visits. International Classification of Diseases codes were referenced against a lookup table that indicated the status and laterality of the clinical conditions of interest[9] (Fig S4, available at www.ophthalmologyscience.org).

NLP Framework

We developed an NLP framework that first extracted text spans indicating DR related clinical concepts from the formatted notes using regular expressions, and then used a suite of machine learning classifiers to infer attributes for the extracted clinical concepts (Fig S5, available at www.ophthalmologyscience.org).

The annotations were consolidated into 7 span-level classification tasks. Using the full set of clinical concepts, we trained 1 classifier to infer status (i.e., whether the condition is present or not present at the time of the visit) and another classifier to infer laterality (i.e., which eye(s) a clinical concept text span refers to). Using concept-specific annotations, we trained 5 additional classifiers to infer severity/type and text span correctness—NPDR severity, PDR severity, RD type, ME type (e.g., DME or other), and ME span correctness. The full set of tasks, their associated clinical concepts, and attribute labels are included in Table S1 (available at www.ophthalmologyscience.org).

Bidirectional Encoder Representations from Transformers-Multilayer Perceptron Classifiers. We trained separate classifiers for each classification task. They leveraged the same classification architecture—a Bidirectional Encoder Representations from Transformers (BERT) encoder[15] with a dense Multilayer Perceptron (MLP) output layer. First, each extracted text span and up to 128 tokens centered around the text span (i.e., the text span's context) were passed through the BERT encoder to generate token-level embeddings. Next, embeddings for tokens from the extracted text span were mean-pooled and concatenated with a one-hot-encoded vector that indicated which clinical concept was represented by the text span. The one-hot vector was included to provide concept-specific priors (biases) on the output. Finally, the concatenated vector representation was passed to the MLP output layer to generate final attribute predictions. A schematic of this classifier architecture (BERT-MLP) is provided in the supplement (Fig S5, available at www.ophthalmologyscience.org).

No existing BERT language model has been trained using clinical text drawn from the ophthalmology domain. We adapted an existing BERT model to our data distribution through a process known as domain adaptive pretraining.[16] We used "Bio_ClinicalBERT"[17] as the initial BERT model, which was previously pretrained on nonclinical web data, PubMed and PubMed Central abstracts, and notes from an intensive care unit setting. We continued pretraining Bio_ClinicalBERT using a masked language modeling objective on text from all ophthalmology provider notes not already included in our annotated dataset.[18]

Training and Validation of NLP Framework

To train and evaluate the task specific classifiers, the annotated dataset was split into 5 mutually exclusive subsets for cross validation—3 were used for training, 1 for model selection (hyperparameter optimization), and 1 to estimate the performance of the model for validation. This process was repeated 5 times, generating 5 unique machine learning classifiers for each task.

Performance of the NLP Classifiers against the gold standard span-level annotations was estimated using macro F1 score (i.e., an unweighted average of class-specific F1 score), precision (i.e., positive predictive value), and recall (i.e., sensitivity). Confidence intervals were estimated using a bootstrap resampling procedure with 100 iterations.[19]

Prediction

We applied the NLP framework to all formatted notes. For each text span, all relevant task specific classifiers of the 35 available (i.e., 7 tasks × 5 training folds) made predictions about the text span's attributes. For each attribute (e.g., laterality, severity), the most frequently predicted label from the 5 associated classifiers was used as the output for subsequent analyses. Ties between classifier predictions were broken randomly. We applied additional postprocessing logic to facilitate our downstream analysis. For RD, we consolidated RD types including TRD and combined rhegmatogenous RD/TRD into RD present. Diabetic macular edema was noted as being present if the ME span was inferred to have a status of present and ME type of DME.

Text-Only NLP System

In the *Text-Only NLP System*, the NLP framework was applied to and made predictions from only the free-text portion of the formatted notes.

Text-and-ICD NLP System

In the *Text-and-ICD NLP System,* the NLP framework was applied to both the free-text and ICD-10 diagnosis codes of the formatted notes. This system used the free-text that surrounds ICD-10 codes in the formatted notes to make probabilistic predictions about the ICD-10 codes' attributes. Importantly, the *Text-and-ICD NLP System* is not a simple additive combination of the outputs of the deterministic *ICD-10 Lookup System* and *Text-Only NLP System*. The *Text-and-ICD NLP System* uses a fully probabilistic approach that leverages text-based context around the ICD-10 codes. The system is able to assign laterality when the deterministic ICD-10 mapping does not specify it, and also differentiates between present (e.g., active) and not present (e.g., negated, resolved) condition statuses. In the example provided in the Supplement, the *Text-and-ICD NLP System* was able to leverage context and correctly identify that the DME, only specified in the ICD-10 code and completely missed by the free-text, was still present in the right eye but had resolved (or was not present) in the left eye (Fig S3, available at www.ophthalmologyscience.org).

Statistical Analyses

Comparing Performance of the Phenotyping Methods with the Gold Standard. The performance of the 3 phenotyping systems for each of the 10 clinical conditions was evaluated at the encounter-level and compared with the gold standard. Macro F1, precision, and recall scores were computed for both per-eye and per-person resolutions. Confidence intervals were estimated using bootstrap resampling with 100 iterations.

Clinical Outcomes

Summary statistics were used to describe the baseline characteristics of patients included in the study (e.g., age, sex, race/ethnicity, insurance). The prevalence of each clinical condition was estimated on the patient level using each system (i.e., *ICD-10 Lookup System*, *Text-Only NLP System*, and *Text-and-ICD NLP System*).

The prevalence estimates as ascertained using the *Text-Only NLP System* or *Text-and-ICD NLP System* were compared with *ICD-10 Lookup System* using the Fisher test. The agreement between the 2 systems for each clinical condition in the per-eye and per-person resolutions was calculated across all encounters using the kappa statistic.[20]

Among office encounters that were identified in both the *Text-Only NLP System* or *Text-and-ICD NLP System* and the *ICD-10 Lookup System*, the date of the earliest encounter in which each clinical condition was identified for each patient was extracted. The difference in the first date of diagnosis as identified by the 2 systems was calculated, with confidence intervals generated using bootstrap resampling with 100 iterations.

All analyses were performed using Python (Python Software Foundation, Python Language Reference, version 3.10.9) and Stata statistical software (version 17.0 for Windows; StataCorp LLC).

Sensitivity Analysis

We performed a sensitivity analysis using only ophthalmology office visits from January 1, 2016 to April 1, 2022. This sensitivity analysis was conducted as ICD-10 was not implemented by health care providers in the United States until October 2015. Cross-mapping of historic ICD-9 to ICD-10 codes at our institution was achieved to variable degrees. Furthermore, evidence shows the possibility of coding inaccuracies in the early use of ICD-10 from October 2015 to January 2016 due to a coding learning curve.[9]

Results

Validation of NLP Framework

Of the 3042 high-recall regular expression matches reviewed by annotators, 3024 (99.4%) were identified as being correctly extracted. On the span-level, the machine learning classifiers achieved average F1, precision, and recall scores (95% confidence interval) of 0.88 (0.83, 0.92), 0.90 (0.88, 0.93), and 0.89 (0.85, 0.93) across the 7 classification tasks. Additional details are provided in the Supplement (Table S2, available at www.ophthalmologyscience.org).

Performance of ICD-10 Lookup System, Text-Only NLP System, and Text-and-ICD NLP System

On the encounter-level, the *Text-and-ICD NLP System* had the highest macro F1 scores across most clinical conditions in both the per-eye and per-person resolutions (Table S3, available at www.ophthalmologyscience.org). The *Text-Only NLP System* performed the best with the highest macro F1 scores for DME, RD, and VH. The *ICD-10 Lookup System* had the highest macro F1 score for severe NPDR in the per-person resolution. Both the *Text-Only NLP System* and *Text-and-ICD NLP System* were able to assign laterality in instances where the *ICD-10 Lookup System* lacked specificity (Tables S4 and S5, available at www.ophthalmologyscience.org).

Comparing ICD-10 Lookup System and Text-Only NLP System for Clinical Outcomes

A total of 91 097 patients and 692 486 office visits were included in the study. Most of the patients were >45 years of age (84%), female (55%), non-Hispanic White (48%), and with Medicare insurance (40%) (Table 6).

In comparing the *ICD-10 Lookup System* and *Text-Only NLP System*, the *Text-Only NLP System* identified a higher prevalence for most clinical conditions (Fig 6). The *ICD-10 Lookup System* identified a higher prevalence of mild NPDR, consolidated NPDR, and NVG. The agreement between the *ICD-10 Lookup System* and *Text-Only NLP System* ranged from slight to moderate (0.03 for unspecified DR to 0.56 for PDR) in the per-eye resolution (Table S7, available at www.ophthalmologyscience.org). Overall, the agreement between the *ICD-10 Lookup System* and *Text-Only NLP System* was higher in the per-person resolution as compared with the per-eye resolution (Table S7, available at www.ophthalmologyscience.org). For all clinical conditions, the *Text-Only NLP System* was able to assign laterality where the *ICD-10 Lookup System* lacked specificity (Tables S8 and S9, available at www.ophthalmologyscience.org). Results were similar on sensitivity analysis (Table S7, available at www.ophthalmologyscience.org).

Table 6. Baseline Characteristics of Patients with Diabetes Mellitus Included in the Study

	Number of Patients (%) N = 91 097 Office Visits: 692 486
Age (yrs)	
≤20	280 (<1%)
>20 to ≤45	13 717 (15%)
>45 to ≤65	40 065 (44%)
>65	37 035 (40%)
Sex	
Male	40 805 (44%)
Female	50 284 (55%)
Other	8 (<1%)
Race/ethnicity	
Non-Hispanic White	43 962 (48%)
Non-Hispanic Black	31 096 (34%)
Hispanic	4358 (5%)
Other	11 681 (13%)
Insurance	
Private	35 873 (39%)
Medicare	36 167 (40%)
Medicaid	8049 (9%)
Other	7671 (8%)
None	1772 (2%)
Missing	1565 (2%)

In comparing across the 3 phenotyping methods, the highest prevalence of nearly all clinical conditions was identified using the *Text-and-ICD NLP System* in the main analysis and sensitivity analysis (Table S7, available at www.ophthalmologyscience.org). The exceptions were unspecified DR in the sensitivity analysis and NVG in both the main and sensitivity analysis (Table S7, available at www.ophthalmologyscience.org, Figs 6 and S7, available at www.ophthalmologyscience.org).

In comparing the date of diagnosis, the *ICD-10 Lookup System* identified an earlier date for nearly all conditions when compared with the *Text-Only NLP System* (Table S7, available at www.ophthalmologyscience.org). The *Text-Only NLP System* identified an earlier year of diagnosis for VH. The first year of diagnosis was similar between the *ICD-10 Lookup System* and *Text-Only NLP System* for severe NPDR and RD. On sensitivity analysis, the *ICD-10 Lookup System* identified an earlier year of diagnosis for all clinical conditions except for VH.

Discussion

Using a BERT-MLP language model, we developed an NLP framework that can accurately identify DR and related conditions (i.e., DR, NPDR, PDR, DME, VH, RD, and NVG) and infer attributes including laterality, status, and severity/type. The NLP framework outperformed the ICD-10 diagnosis codes-only approach in F1 score, precision,

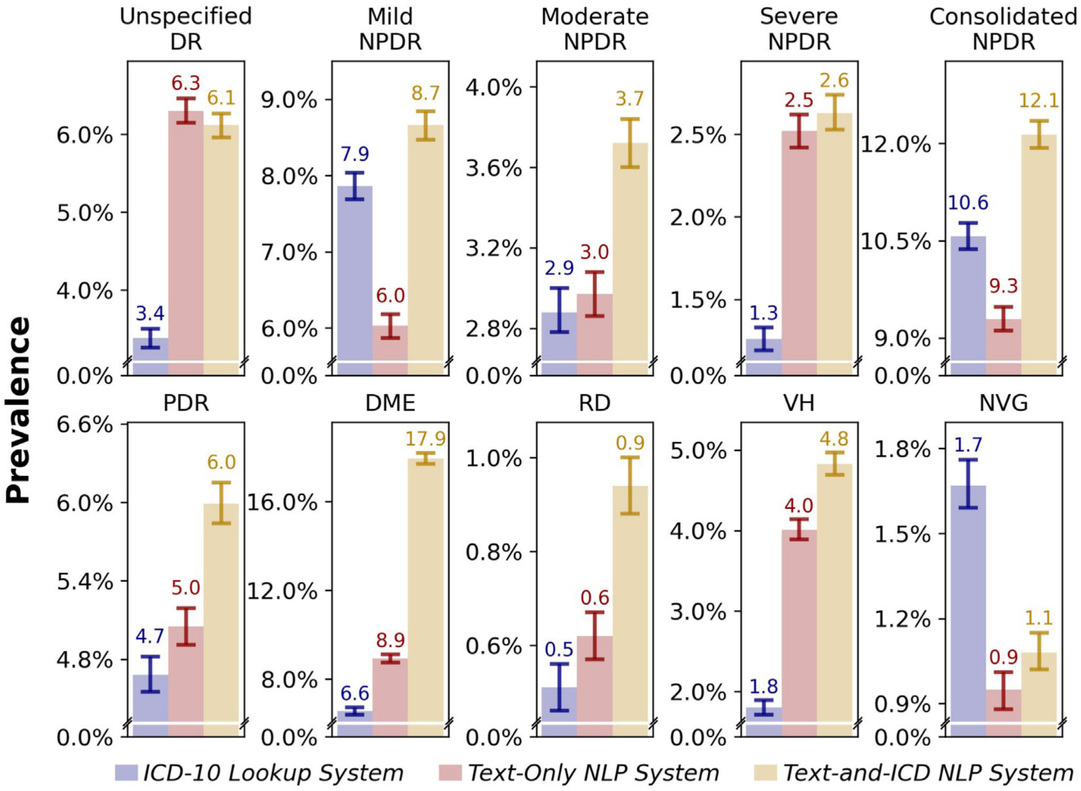

Figure 6. The percent prevalence (and 95% confidence interval) of DR and related conditions as identified by each of the 3 phenotyping methods (*ICD-10 Lookup System*, *Text-Only NLP System*, *Text-and-ICD NLP System*). DME = diabetic macular edema; DR = diabetic retinopathy; ICD-10 = International Classification of Diseases, 10th Revision; NLP = natural language processing; NPDR = nonproliferative diabetic retinopathy; NVG = neovascular glaucoma; PDR = proliferative diabetic retinopathy; RD = retinal detachment; VH = vitreous hemorrhage.

and recall for each of the NLP classification tasks. Compared with ICD diagnosis codes alone, we were able to identify more cases, or a higher prevalence, of most DR related conditions using the provider free-text notes; however, this varied substantially depending on the clinical condition. We achieved the best performance using our combined *Text-and-ICD NLP system* that made context-informed probabilistic predictions about clinical concepts indicated by ICD-10 diagnosis codes and within the free-text. The combined system generally had the highest F1 score and identified the highest prevalence of nearly all DR and related conditions.

Natural language processing is increasingly used in ophthalmology, particularly for the identification of clinical conditions.[21–23] Multiple studies have shown that case selection is greatly enhanced with the application of NLP to free-text notes, and that diagnosis codes are either inadequate to capture all cases of a disease of interest, or only capture a small portion of those cases.[21,22,24] We similarly show that for most DR and related conditions, up to twofold to threefold more cases could be identified using free-text notes but this was not true for all conditions. The variability in the clinical condition prevalences identified across our 3 phenotyping systems highlights the impact of local coding practices and documentation workflow on case identification using ICD-10 codes alone. For example, for conditions such as VH that are poorly coded,[9] application of the NLP to free-text was superior when compared with ICD-10 codes alone (identifying a prevalence of 4.01% compared to 1.80%). However, for conditions such as NVG, more cases were identified using ICD-10 codes as compared with NLP of free-text (1.67% compared to 0.95%). Despite identifying more cases, ICD-10 codes alone had lower precision and recall as compared with the free-text. International Classification of Diseases, 10th Revision codes appeared to be identifying NVG cases that had already been resolved. Similarly with mild NPDR, ICD-10 codes identified a higher prevalence with a higher recall compared with NLP of free-text but at the expense of precision.

A unique aspect of our study was that we were able to directly compare the accuracy of each phenotyping system compared with the gold standard chart review. In circumstances where such chart review is not possible and one cannot critically examine the impact of local coding patterns on accuracy, it is even more advantageous to use a combined system that uses both free-text as well as diagnosis codes, as in our *Text-and-ICD NLP System*. We designed our NLP framework as an aggregation of 7 classifiers rather than a single multilabel classification based on preliminary experiments not reported in this paper. We experimented with a multitask learning setup in which a single BERT encoder was used with multiple classification heads for each attribute type. Throughout these experiments, performance either matched that of the independent models, or fell below that of the independent models, possibly because of negative transfer. Furthermore, due to the significant sample size variation across attributes, we found that the checkpoint in the training process at which an "optimal" level of performance for a particular attribute was achieved frequently differed from the optimal checkpoint of other attributes. In this case, we would still have needed multiple versions of the same model to achieve optimal levels of performance across all tasks.

There are other advantages to our NLP framework. A major advantage is its ability to infer the laterality of the DR related condition. The NLP framework was able to identify laterality in all relevant encounters with excellent performance. The ability of the NLP framework to assign laterality is why the per-eye agreement with the ICD-10 only system is lower than that of the per-person agreement. Another advantage is that it is able to infer status, present or not present, again with excellent performance. Furthermore, the NLP framework is able to assign severity to DR, for example, mild/moderate/severe in cases where the ICD-10 only had unspecified DR. This is why the identified prevalence of unspecified DR by the NLP framework is lower than that of the ICD-10 only system.

In the context of DR, prior NLP efforts to identify the stage of DR have leveraged MediClass, a knowledge-based system that detects clinical events.[25,26] The NLP system developed here leverages the same 2-stage structure employed by MediClass, namely concept identification and classification. In both cases, concept identification is facilitated using a domain-specific knowledge base which contains relevant clinical concepts/events and sets of terms/phrases that represent them. While there are minor implementation differences in how the 2 systems extract concepts from the free-text, the most significant difference lies in the classification stage. Whereas MediClass requires manual specification of logical rules to validate the correctness of a span extraction and to assign appropriate attributes (e.g., laterality, severity), our NLP framework uses a contextual language model and statistical classifier which automatically *learns* linguistic relationships using labeled examples to perform the same tasks. A recent study adopted a similar approach to our NLP framework, but it was only designed to operate on imaging reports from patients already diagnosed with DR, and was unable to infer attributes of clinical concepts if they were not explicitly mentioned in the free-text.[27]

In comparing the date of diagnosis, the *ICD-10 Lookup System* often identified an earlier date of onset compared with the *Text-Only NLP System* but this difference was generally limited to a few months. This systematic difference could reflect coding patterns unique to our institution, where optometrists who often make an initial DR related diagnosis are more likely to use problem based documentation before referring to a retina specialist who is more likely to use progress notes. One notable exception is VH where NLP of the free-text often identified the clinical condition much earlier than ICD-10 codes by on average 4 months. This is again consistent with prior findings that VH, a condition that is intermittent and often self-resolving, is poorly coded.[9]

We identify a lower prevalence of milder disease and a higher prevalence of more severe and vision-threatening DR compared with national estimates. Nationally, the prevalence of mild NPDR ranges from 16.3% to 36.6%, moderate NPDR 1.7% to 10.3%, severe NPDR or PDR 1.0% to 6.9%, and DME 1.2% to 8.9%.[28] In our study, we identify a lower prevalence of the milder diseases including mild NPDR (8.7%), moderate NPDR (3.7%), and slightly higher prevalence of vision-threatening DR including PDR (6.0%) and DME (17.9%). When defining vision threatening DR to include severe NPDR, PDR, or DME, we find a prevalence of 18.6% (using the *Text-and-ICD NLP System*), which is higher than previous reports of 5.1% to 8.2%.[5,28] This is not surprising as tertiary care referral centers tend to have more severe disease.

There are fewer studies characterizing the prevalence of other neovascular sequelae of DR including VH, TRD, and NVG. A prospective study of patients presenting with first time spontaneous VH estimates an incidence of 7 cases per 100 000 inhabitants.[29] Another study using administrative claims estimates a crude incidence rate of 4.8 cases per 10 000 person-years.[30] Among patients with PDR enrolled in a prospective clinical trial, 5.6% developed VH or TRD requiring surgical intervention with pars plana vitrectomy.[31] In a study using administrative claims in Optum, Gange et al quote an incidence of TRD of 0.3%, and NVG 0.1%.[6] It is difficult to compare the estimates for these DR related conditions to those in this study as they are slightly different measures (incidence compared to prevalence). In general, we find a prevalence of VH of 4.8%, 0.9% for RD (which combines TRD and TRD/rhegmatogenous RD), and 1.1% for NVG.

There are limitations to this work. One disadvantage of the NLP framework is its performance with respect to rare events in the annotation dataset (e.g., non−high-risk PDR and incorrect ME text spans). Future dataset expansion efforts may benefit from active learning techniques to more efficiently annotate rare targets.[27,32] We also have not quantified the degree to which the initial list of high-recall regular expressions could be missing relevant concepts. Based on domain expert knowledge, we suspect that the false negative rate is low. However, since the NLP system was developed at a single institution, we do not know the degree to which our algorithm is generalizable to other institutions. This could also limit the scalability of the framework. Another disadvantage of the NLP framework is its computational expense requirements. Due in part to these computational limitations, we only explored the BERT-MLP architecture with a handful of possible encoder initialization strategies. Alternative architectures (e.g., generative pretrained transformers and text-to-text transfer transformer) may have yielded better performance, though we note that BERT models continue to achieve state-of-the-art levels of performance in a variety of clinical and biomedical tasks when annotated data is available.[33−35] Furthermore, as described in our methods-focused study,[14] adaptation of the language model to our clinical language domain via continued pretraining and task fine tuning provides a benefit to performance well beyond what is provided by choosing a different encoder initialization. When training the classifiers, we did not include examples in which the attribute label could not be determined by the annotators (i.e., unspecified). This exclusion during training means it is possible that the classifier's performance estimate could be overconfident, with the examples containing specified attributes potentially being easier. Although we use a random sample of patients from our population, there is a possibility that our system performs differently for subgroups within the population (e.g., performs worse for female patients or certain race and ethnicity groups or performs worse for notes written by providers with a certain level of training). While the training sample is still representative of the target population, any group imbalance in the training data can implicitly motivate the model to perform better on majority groups.[36] Lastly, although we discuss the prevalence of disease, we recognize that since these values were derived from a single institution, they do not represent national prevalence. Despite these limitations, we share a useful NLP framework that can identify DR and related conditions and infer attributes including laterality, status, and severity/type. Future work using our NLP framework should focus on establishing the external validity of the system and its performance at other institutions. This system will enable more detailed analyses of EHR data at scale to derive clinically meaningful insights around DR.

In conclusion, we provide a comprehensive comparison between ICD-10 diagnosis codes and free-text note identification of many crucial concepts in the study of DR and related conditions, showing the strengths and weaknesses of each approach. The best performing phenotyping method was the *Text-and-ICD NLP System* that used information in both diagnosis codes as well as free-text notes.

Footnotes and Disclosures

Originally received: April 19, 2024.
Final revision: July 4, 2024.
Accepted: July 12, 2024.
Available online: July 18, 2024. Manuscript no. XOPS-D-24-00120.

[1] Department of Computer Science, Johns Hopkins University, Baltimore, Maryland.

[2] Wilmer Eye Institute, Johns Hopkins School of Medicine, Baltimore, Maryland.

[3] Department of Biomedical Informatics and Data Science, Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, Maryland.

[4] Center for Population Health Information Technology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland.

Machine Learning for Health (ML4H), New Orleans, LA. December 10, 2023.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have made the following disclosures:

M.D.: Consultant – Bloomberg LP; Equity – Good Analytics.

C.X.C.: Grants – Regeneron.

This project was supported by a Career Development Award from the Research to Prevent Blindness (C.X.C.), K23 award from the NIH/NEI (award number K23EY033440) (C.X.C.), and an unrestricted grant from Research to Prevent Blindness (Wilmer Eye Institute). C.X.C. is the Jonathan and Marcia Javitt Rising Professor of Ophthalmology.

HUMAN SUBJECTS: Human subjects were included in this study. The study was approved by the Johns Hopkins Institutional Review Board which waived the requirement for patient consent as this was secondary research. All research adhered to the tenets of the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Harrigian, Dredze, Cai

Data collection: Harrigian, Tran, Tang, Gonzales

Analysis and interpretation: Harrigian, Nagy, Kharrazi, Dredze, Cai

Obtained funding: Cai

Overall responsibility: Harrigian, Dredze, Cai

Abbreviations and Acronyms:
BERT = Bidirectional Encoder Representations from Transformers; **DME** = diabetic macular edema; **DR** = diabetic retinopathy; **EHR** = electronic health record; **ICD** = International Classification of Diseases; **ICD-10** = International Classification of Diseases, 10th Revision; **ME** = macular edema; **MLP** = Multilayer Perceptron; **NLP** = natural language processing; **NPDR** = nonproliferative diabetic retinopathy; **NVG** = neovascular glaucoma; **PDR** = proliferative diabetic retinopathy; **RD** = retinal detachment; **TRD** = tractional retinal detachment; **VH** = vitreous hemorrhage.

Keywords:
Clinical free-text notes, Diabetic retinopathy, Electronic health record, Natural language processing, Prevalence.

Correspondence:
Cindy X. Cai, MD, MS, Wilmer Eye Institute, 1800 Orleans Street, Room 711, Baltimore, MD 21287. E-mail: ccai6@jhmi.edu.

References

1. Klein R, Klein BEK. *Vision disorders in diabetes. Diabetes in America*. 2nd ed. National Institutes of Health; 1995:293–338.
2. Saaddine JB, Honeycutt AA, Narayan KMV, et al. Projection of diabetic retinopathy and other major eye diseases among people with diabetes mellitus: United States, 2005-2050. *Arch Ophthalmol*. 2008;126(12):1740–1747.
3. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med*. 2015;12(10): e1001885.
4. Big data analytics for personalized medicine. *Curr Opin Biotechnol*. 2019;58:161–167.

5. Lundeen EA, Burke-Conte Z, Rein DB, et al. Prevalence of diabetic retinopathy in the US in 2021. *JAMA Ophthalmol*. 2023;141(8):747−754.
6. Gange WS, Lopez J, Xu BY, et al. Incidence of proliferative diabetic retinopathy and other neovascular sequelae at 5 Years following diagnosis of type 2 diabetes. *Diabetes Care*. 2021;44(11):2518−2526.
7. Lau M, Prenner JL, Brucker AJ, VanderBeek BL. Accuracy of billing codes used in the therapeutic care of diabetic retinopathy. *JAMA Ophthalmol*. 2017;135(7):791−794.
8. Muir KW, Gupta C, Gill P, Stein JD. Accuracy of international classification of diseases, ninth revision, clinical modification billing codes for common ophthalmic conditions. *JAMA Ophthalmol*. 2013;131(1):119−120.
9. 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(4):374−380.
10. Hwang TS, Thomas M, Hribar M, et al. The impact of documentation workflow on the accuracy of the coded diagnoses in the electronic health record. *Ophthalmol Sci*. 2024;4(1): 100409.
11. Hatfield M, Nguyen TH, Chapman R, et al. Identifying the mechanism of missingness for unspecified diabetic retinopathy disease severity in the electronic health record: an IRIS® Registry analysis. *J Am Med Inform Assoc*. 2023;30(6): 1199−1204.
12. Cai CX, Tran D, Tang T, et al. Health disparities in lapses in diabetic retinopathy care. *Ophthalmol Sci*. 2023;3(3):100295.
13. GitHub - kharrigian/ml4h-clinical-bert: Official repository for the ML4H (Findings) paper: “An Eye on Clinical BERT: Investigating Language Model Generalization for Diabetic Eye Disease Phenotyping.” GitHub. https://github.com/kharrigian/ml4h-clinical-bert. Accessed March 6, 2024.
14. Harrigian K, Tang T, Gonzales A, et al. An eye on clinical BERT: investigating language model generalization for diabetic eye disease phenotyping. http://arxiv.org/abs/2311.08687; 2023. Accessed December 27, 2023.
15. Jacob D, Chang MW, Lee K, Toutanova K. BERT: pretraining of deep bidirectional Transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics; 2019:4171−4186. https://aclanthology.org/N19-1423.pdf. Accessed January 1, 2024.
16. Gururangan S, Marasović A, Swayamdipta S, et al. Don’t stop pretraining: adapt language models to domains and tasks. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2020:8342−8360. Online https://aclanthology.org/2020.acl-main.740.pdf. Accessed January 1, 2024.
17. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019:72−78. https://aclanthology.org/W19-1909.pdf. Accessed January 1, 2024.
18. Garrett Wilson Washington State University Pullman. Pullman, WA, Washington, Diane J. Cook Washington state University Pullman, Pullman, WA, Washington. A Survey of unsupervised deep domain adaptation. *ACM Trans Intell Syst Technol*. 2020;11:1−46.
19. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Stat Sci*. 1996;11(3):189−228.
20. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37−46.
21. 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.
22. Baxter SL, Klie AR, Radha SB, et al. Text processing for detection of fungal ocular involvement in critical care patients: cross-sectional study. *J Med Internet Res*. 2020;22(8): e18855.
23. Chen JS, Baxter SL. Applications of natural language processing in ophthalmology: present and future. *Front Med*. 2022;9:906554.
24. Maganti N, Tan H, Niziol LM, et al. Natural Language processing to quantify microbial keratitis measurements. *Ophthalmology*. 2019;126(12):1722−1724.
25. Smith DH, Johnson ES, Russell A, et al. Lower visual acuity predicts worse utility values among patients with type 2 diabetes. *Qual Life Res*. 2008;17(10):1277−1284.
26. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: a system for detecting and classifying encounter-based clinical events in any electronic medical record. *J Am Med Inform Assoc*. 2005;12(5):517−529.
27. Cohn DA, Ghahramani Z, Jordan MI. Active learning with statistical models. *JAIR*. 1996;4:129−145. arXiv:cs/9603104.
28. Kempen JH, O’Colmain BJ, Leske MC, et al. The prevalence of diabetic retinopathy among adults in the United States. *Arch Ophthalmol*. 2004;122(4):552−563.
29. Lindgren G, Sjödell L, Lindblom B. A prospective study of dense spontaneous vitreous hemorrhage. *Am J Ophthalmol*. 1995;119(4):458−465.
30. Wang CY, Cheang WM, Hwang DK, Lin CH. Vitreous haemorrhage: a population-based study of the incidence and risk factors in Taiwan. *Int J Ophthalmol*. 2017;10(3): 461−466.
31. Flynn Jr HW, Chew EY, Simons BD, et al. Pars plana vitrectomy in the early treatment diabetic retinopathy study. ETDRS report number 17. The early treatment diabetic retinopathy study research group. *Ophthalmology*. 1992;99(9): 1351−1357.
32. Huang PY. A Survey of Deep Active Learning. *ACM Comput Surv*. 2021;54:1−40.
33. Lehman E, Hernandez E, Mahajan D, et al. Do we still need clinical language models? In: *Proceedings of the Conference on Health, Inference, and Learning,* Proceedings of Machine Learning Research, 2023:578−597, 209. https://proceedings.mlr.press/v209/eric23a.html. Accessed August 6, 2024.
34. Labrak Y, Rouvier M, Dufour R. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2307.12114. preprint arXiv:2307.12114.
35. Rehana H, Çam NB, Basmaci M, Zheng J, Jemiyo C, He Y, Hur J. Evaluation of GPT and BERT-based models on identifying protein-protein interactions in biomedical text. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2303.17728. preprint arXiv:2303.17728.
36. Sagawa S, Raghunathan A, Koh PW, Liang P. An investigation of why overparameterization exacerbates spurious correlations. In: *Proceedings of the 37th International Conference on Machine Learning,* PMLR; 2020:119. Online http://proceedings.mlr.press/v119/sagawa20a/sagawa20a.pdf. Accessed January 1, 2024.