AMERICAN ACADEMY OF OPHTHALMOLOGY®

Automated Recognition of Visual Acuity Measurements in Ophthalmology Clinical Notes Using Deep Learning

Isaac A. Bernstein, BS,[1] *Abigail Koornwinder, BS,*[1] *Hannah H. Hwang, BS,*[2] *Sophia Y. Wang, MD, MS*[1]

Purpose: Visual acuity (VA) is a critical component of the eye examination but is often only documented in electronic health records (EHRs) as unstructured free-text notes, making it challenging to use in research. This study aimed to improve on existing rule-based algorithms by developing and evaluating deep learning models to perform named entity recognition of different types of VA measurements and their lateralities from free-text ophthalmology notes: VA for each of the right and left eyes, with and without glasses correction, and with and without pinhole.

Design: Cross-sectional study.

Subjects: A total of 319 756 clinical notes with documented VA measurements from approximately 90 000 patients were included.

Methods: The notes were split into train, validation, and test sets. Bidirectional Encoder Representations from Transformers (BERT) models were fine-tuned to identify VA measurements from the progress notes and included BERT models pretrained on biomedical literature (BioBERT), critical care EHR notes (ClinicalBERT), both (BlueBERT), and a lighter version of BERT with 40% fewer parameters (DistilBERT). A baseline rule-based algorithm was created to recognize the same VA entities to compare against BERT models.

Main Outcome Measures: Model performance was evaluated on a held-out test set using microaveraged precision, recall, and F1 score for all entities.

Results: On the human-annotated subset, BlueBERT achieved the best microaveraged F1 score (F1 = 0.92), followed by ClinicalBERT (F1 = 0.91), DistilBERT (F1 = 0.90), BioBERT (F1 = 0.84), and the baseline model (F1 = 0.83). Common errors included labeling VA in sections outside of the examination portion of the note, difficulties labeling current VA alongside a series of past VAs, and missing nonnumeric VAs.

Conclusions: This study demonstrates that deep learning models are capable of identifying VA measurements from free-text ophthalmology notes with high precision and recall, achieving significant performance improvements over a rule-based algorithm. The ability to recognize VA from free-text notes would enable a more detailed characterization of ophthalmology patient cohorts and enhance the development of models to predict ophthalmology outcomes.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100371*

Supplemental material available at www.ophthalmologyscience.org.

More information than ever is stored in free-text notes within the electronic health record (EHR), including detailed descriptions of patients' symptoms, physical examination, and the physician's assessment and plan. Especially critical in the field of ophthalmology is the eye examination portion of clinical notes, which includes an assessment of visual acuity (VA). Visual acuity gauges a patient's ability to recognize different figures or "optotypes" at a standard distance, and, in doing so, evaluates refraction, retina function, and higher-order cognitive processing to interpret visual stimuli.[1] Visual acuity is critical in the study of ophthalmic disease because it allows detailed characterization of patient phenotypes, the definition of patient cohorts, and greater understanding of treatment trajectories. It may also be used to develop predictive models for patient outcomes, both as an input feature and as an important patient outcome. Visual acuity is clinical information that is only available in the health record. Although some EHR software includes semistructured data fields designed to facilitate easy input and extraction of ophthalmic variables, not all systems include this feature; in those cases, even the most important clinical information, such as VA, may be sequestered as unstructured free-text notes, requiring natural language processing techniques to process, understand, and compute over.

https://doi.org/10.1016/j.xops.2023.100371
ISSN 2666-9145/23

Transformer-based deep learning architectures such as Bidirectional Encoder Representations from Transformers (BERT) and its descendants have achieved high performance on many natural language processing tasks, including named entity recognition (NER) tasks.[2] Bidirectional Encoder Representations from Transformers models trained on massive corpora of internet text are publicly available, as well as extensions of BERT further trained on biomedical domain text. These models include BioBERT[3] (pretrained on biomedical literature [PubMed]), ClinicalBERT[4] (pretrained on critical care EHR notes [MIMIC-III]), and BlueBERT[5] (pretrained on both biomedical literature and critical care EHR notes). DistilBERT[6] is a smaller version of BERT with 40% fewer parameters. Despite not having been pretrained on biomedical domain-specific texts, DistilBERT has been shown to perform comparably to ClinicalBERT and BioBERT on biomedical NER tasks, including recognition of protected health information.[7] DistilBERT's advantage is that, owing to its relatively smaller size, it is less computationally expensive to train and perform inference with.

One major limitation of using deep learning models is the need for large sets of high-quality training data. These data sets are especially difficult, time-consuming, and expensive to create in medical fields, such as ophthalmology, in which domain expertise is required and data sets consist of protected health information. Weak supervision is an approach that can be used to efficiently create large amounts of training data by leveraging data structures, patterns, rules, or other classifiers to label the corpora in an automated fashion, enabling the training of deep learning models for NER that would be otherwise infeasible.[8] Here, we aimed to build and evaluate deep learning models that can identify VA measurements and their type and laterality from unstructured, free-text ophthalmology notes. These models are fine-tuned for this task using a large data set created using a weakly supervised approach, and their application for identifying VA from free-text notes does not require pre-existing VA-specific data structures in the EHR. Applying deep learning-based recognition of VA to the ever-increasing amounts of unstructured data available in EHR systems could catalyze the study of ophthalmic disease and the development of clinical tools at scale.

Methods

Overview

We compared the performance of models initialized on pretrained DistilBERT, BioBERT, ClinicalBERT, and BlueBERT on the task of recognizing named entities documenting VA in ophthalmology clinical notes. The HuggingFace library[9] was used to fine-tune pretrained BERT models on our data, which consisted of ophthalmology clinical progress notes labeled with 8 different types of VA measurements.

Data Source

We identified from the Stanford Research Repository all of the clinical notes and VA measurements of patients who were seen by the Department of Ophthalmology at Stanford University since 2008,[10] documented on a single EHR system (Epic Systems Corporation). From the total of 333 958 notes belonging to approximately 90 000 patients, notes missing corresponding VA labels were excluded, resulting in a final sample of 319 756 notes for the study. Data were split into train, validation, and test sets at a ratio of 80:10:10. Full notes were split into shorter subdocuments to accommodate input into models, which have specified maximum length of tokens: 512 for both DistilBERT and BioBERT and 128 for both BlueBERT and ClinicalBERT. This study was approved by the Stanford University institutional review board. The institutional review board granted a waiver of informed consent given the scale of the data and observational nature of the data set. The study adhered to the tenets of the Declaration of Helsinki.

Preprocessing Labels

Technicians at this institution use semistructured fields to report VA measurements in the eye examination portion of the EHR system. There are 8 different classes of VA measurements (named entities) that we sought to identify, including VA for each of the right (OD) and left eyes (OS), with and without glasses correction, and with and without pinhole. Once the technician enters the VA into the semi-structured field, the information is usually imported into the clinical free-text note using the providers' custom note templates. Thus, the semistructured fields were VA labels for their corresponding free-text clinical notes, which were the inputs to the models.

Figure 1 illustrates a semistructured field with its corresponding clinical progress note. Visual acuity labels are known at the document level rather than the token level; from information entered into the semistructured field, we know what the VA of the patient is, but we do not necessarily know which exact words or tokens in the note correspond to that VA. Therefore, a custom preprocessing pipeline was developed to assign token-level labels for each document ("training labels"). The full preprocessing pipeline is illustrated in Figure 2.

Each document was pretokenized using the Treebank Tokenizer in the Python Natural Language ToolKit version 3.5.[11] For each documented VA measurement in the semistructured field, we found which tokens corresponded to that measurement in the note. In some cases, there could be multiple matches, such as in cases in which the OD and OS VAs were the same. A greedy process was used to assign each label to a token, iterating through each VA label and assigning the first matched token to that label if the token was not already assigned to another VA label. This makes use of the assumption that, in most cases, reporting of VA starts with the right side. Labels were constructed in the Inside-Outside-Beginning format,[12] with "O" for no label or outside of the entity, "B-valabel" for the beginning of a VA entity, and "I-valabel" for tokens that continue (or are inside of) a VA entity. The result of this process is a list of tokens and a list of corresponding VA labels.

Bidirectional Encoder Representations from Transformers word piece tokenization was performed to further break down tokens into word pieces.[13] Original full word token-level labels were then "propagated" as appropriate to each word piece token: "O" labels were assigned to each word piece within a word labeled with "O," and "I" labels were assigned to each word piece within a word labeled "B" or "I." Padding and truncation were used to standardize the length of each subdocument for input into models.

Evaluation on Human-Annotated Ground Truth Set

The weak supervision approach involved training models on semiautomatically labeled data using semistructured fields, which may have generated errors in labels. Therefore, we further

evaluated model performance on a sample of 300 manually labeled ophthalmology notes from the test set. We used the Prodigy annotation tool[14] to visualize, analyze, and correct our models' predictions. From the annotated sample of documents, we report the standard evaluation metrics of each model (and its training labels) to the ground truth human annotation. We also give qualitative examples of typical contexts in which the models fail.

Modeling and Experimental Details

All pretrained models were initialized through the HuggingFace transformers library for the token classification task and fine-tuned on our task to identify the 8 different types of VA named entities. All models were trained with standard cross-entropy loss function for token classification, with the Adam optimizer (learning rate 5e−5 and weight decay of 0.01) and warmup steps of 500. Early stopping was used with patience equaling 3. Validation loss was calculated after each epoch, and the model with the best validation loss was then used for final evaluation. The HuggingFace library was used with the Optuna backend[15] for hyperparameter tuning of the trained models of the validation set. Hyperparameters optimized included learning rate, weight decay, number of epochs to train for, and number of warmup steps. Hyperparameter search was initialized for 5 trials seeking to minimize validation loss. The best-identified hyperparameters (Table S1, available at www.ophthalmologyscience.org) were used to retrain models, this time without early stopping, and final evaluation of the models was performed on the test set.

Baseline Model

For our baseline classifier, we developed a rule-based algorithm using regular expressions (regex) to identify VA-based keywords and abbreviations corresponding to with or without correction, with and without pinhole. The regexes used are presented in Table S2, available at www.ophthalmologyscience.org. Regex 1 through 4 identified these keywords related to correction in the distance VA examination section and their corresponding VAs. The algorithm then checked if both VAs were presented consecutively after a keyword or if only one VA was attached to each keyword. In the consecutive case, the first VA was assigned to OD, and the second was assigned to OS. If only one VA is attached to a keyword, the first occurrence of a keyword and VA was assigned to the right, and the second occurrence, if any, was assigned to the left. We followed this assignment logic because the OD acuity was reported first in almost all cases. Regex 5 was then used to extract VA from its corresponding keyword. The algorithm then checked if additional VAs were reported with pinhole (regex 6 through 8), and appropriate laterality was assigned. Additional regex were used to flag multiword nonnumeric VAs (e.g., count fingers at 3', hand motion at 3') so that the algorithm could properly assign the first word in the VA with a beginning tag (B-) and the subsequent words with an inside tag (I-), following Inside-Outside-Beginning format. The general flowchart of this algorithm is outlined in Figure S3, available at www.ophthalmologyscience.org.

Evaluation Metrics

The performance of our NER system was evaluated using the Python seqeval package (version 1.2.2),[16] a framework for sequence labeling evaluation. We report precision (number of correctly predicted entities/total number of predicted entities), recall (number of correctly predicted entities divided by total number of labeled entities), and F1 score (2 × precision × recall divided by [precision + recall]) for each named entity class. Microaveraged metrics were also computed across all classes, defined as the sum of correctly predicted entities for all the classes divided by the sum of the total number of predicted entities for all the classes.

Code Availability

Code associated with this project is publicly available.[17]

Results

Common Sources of Training Label Errors

We compared the training labels and each of the model predictions with the human-annotated ground truth on a sample of 300 inputs from the test set. Because the training labels were algorithmically derived, performing human review of the training labels gives a sense of the magnitude of the noise in the training labels (Table S3, available at www.ophthalmologyscience.org). Overall, training labels were very close to human-annotated ground truth, with a microaveraged F1 score of 0.87. From qualitative review during the human annotation of the training labels, it was noted that VAs were sometimes not labeled in the examination portion but, instead, labeled in other sections of the note, such as the refraction or assessment and plan sections. However, VAs were still correctly recognized. Examples of this error are presented in Figure S1, available at www.ophthalmologyscience.org.

Model Performance Against Training Labels

Metrics assessing the performance of each fine-tuned model on the held-out test set for each VA type are shown in Table 1. Overall, BioBERT achieved the best microaveraged F1 score at 0.90, followed by BlueBERT (0.89), DistilBERT (0.87), the baseline model (0.76), and ClinicalBERT (0.75). BioBERT also had the highest microaveraged precision (0.89) and recall (0.91). The best-performing VA type for this model was OS, without correction, with pinhole, with an F1 score of 0.92, precision of 0.93, and recall of 0.91. BioBERT's worst performing VA type was OS, without correction, no pinhole with an F1 score of 0.88, precision of 0.85, and recall of 0.91.

Model Performance Against Human-Annotated Ground Truth

The performance metrics for each model's predictions compared with the human-annotated ground truth are shown in Table 2. For these metrics, the biomedical models had comparable microaveraged F1 scores, with BlueBERT being the best (0.92), followed by ClinicalBERT (0.91), DistilBERT (0.90), BioBERT (0.84), and, lastly, the baseline model (0.83). Whereas BlueBERT held the highest recall (0.97), the baseline model had the highest precision (0.93) but the lowest recall (0.76). Examples of common model errors are given in Figure S2. Common mistakes included missing nonnumeric VA values, such as no improvement, no light perception, count fingers, and hand motion, recognizing VAs in the wrong portion of the note, such as in the refraction or assessment and plan section, difficulties recognizing VA along a series of visual

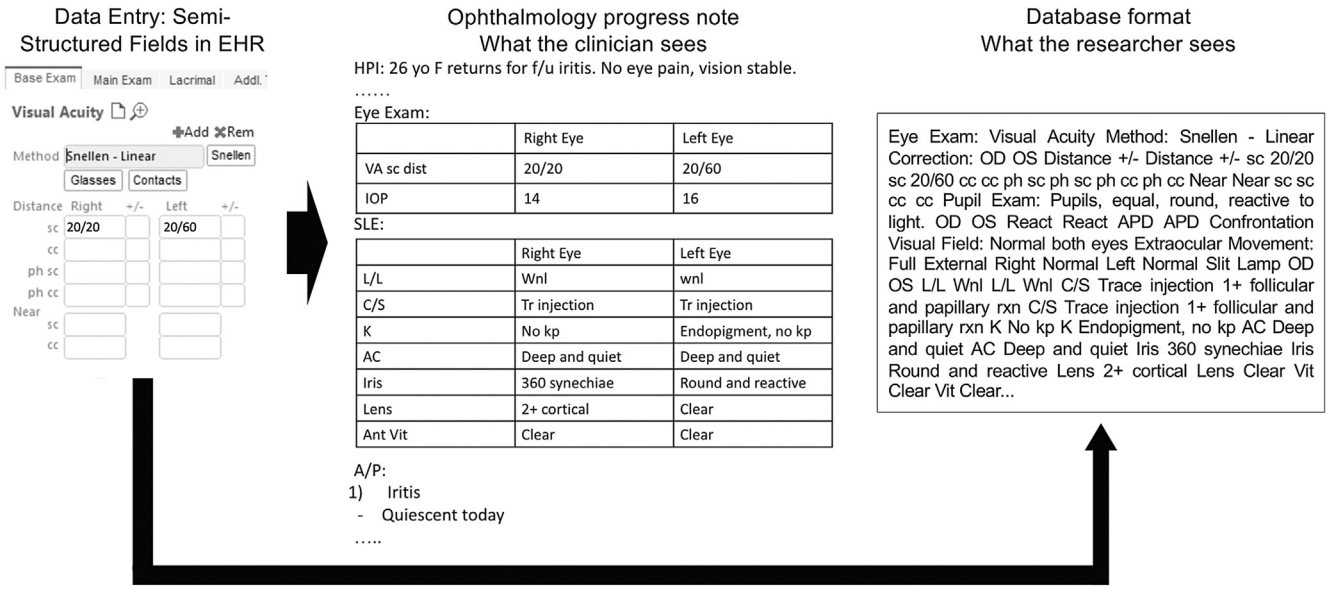

Figure 1. Electronic health records (EHRs) free text data format.

acuities (depicting VA changes over several dates, for instance) misidentifying OD as OS and vice-versa, only recognizing the numerator or denominator of the measured VA, and recognizing "without correction" as CC instead of SC. Figure S3-S6, available at www.ophthalmologyscience.org, additionally highlight instances in which each type of model was correct while others failed, and Figure S7, available at www.ophthalmologyscience.org, compares examples between the baseline model and the BERT models. There was a lack of evident patterns that clearly differentiated between the different types of BERT models. However, there were consistent instances in which the regular expression model performed poorly compared with BERT models, likely due to the limited adaptability of regular expressions (Tables S1, S2 and S4, available at www.ophthalmologyscience.org).

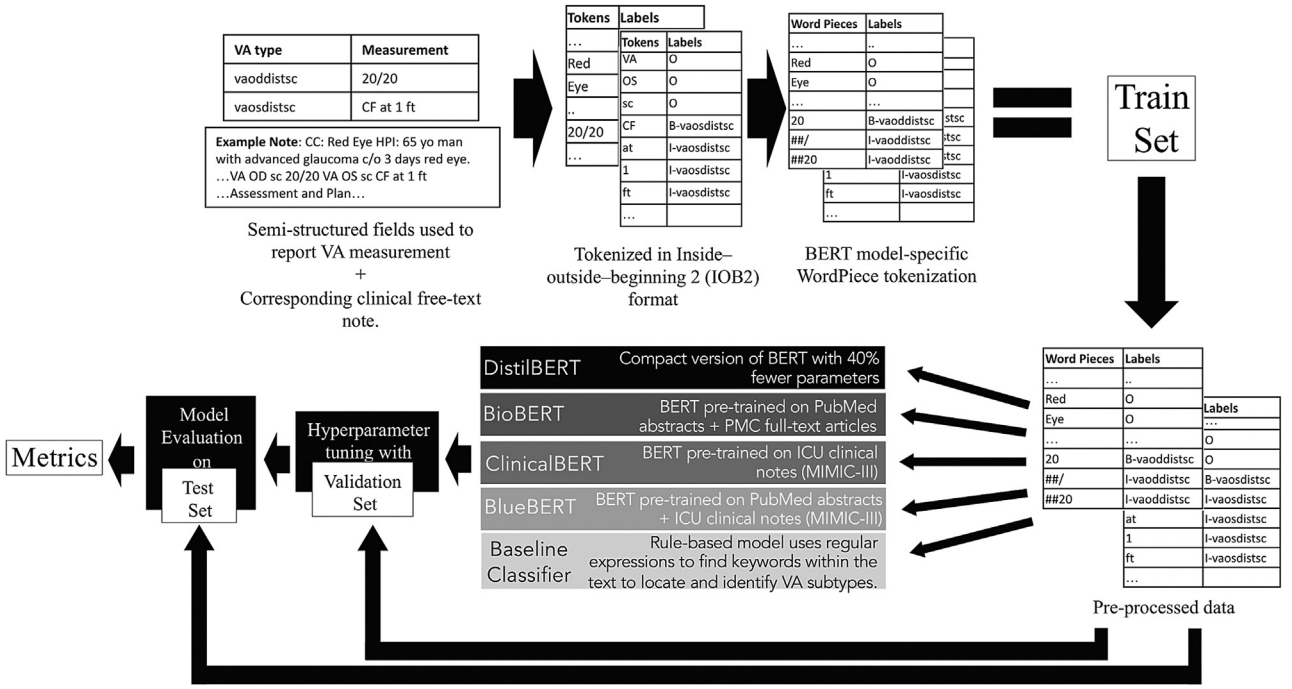

Figure 2. Pre-processing pipeline.

Table 1. Model Performance on Test Set

		OD				OS				
		CC		SC		CC		SC		
Model	**Metric**	*No PH*	*PH*	*No PH*	*PH*	*No PH*	*PH*	*No PH*	*PH*	**Microaveraged**
Baseline model	Precision	0.87	0.60	0.85	0.70	0.92	0.90	0.93	0.93	0.85
	Recall	0.74	0.82	0.74	0.86	0.66	0.44	0.67	0.51	0.69
	F1	0.80	0.69	0.79	0.77	0.77	0.59	0.78	0.66	0.76
DistilBERT	Precision	0.83	0.85	0.84	0.90	0.79	0.89	0.83	0.92	0.84
	Recall	0.93	0.94	0.88	0.93	0.93	0.91	0.86	0.94	0.91
	F1	0.87	0.89	0.86	0.91	0.86	0.90	0.85	0.93	0.87
BioBERT	Precision	0.93	0.87	0.87	0.88	0.91	0.88	0.85	0.93	0.89
	Recall	0.89	0.94	0.95	0.95	0.87	0.93	0.91	0.91	0.91
	F1	0.91	0.90	0.91	0.91	0.89	0.90	0.88	0.92	0.90
ClinicalBERT	Precision	0.64	0.74	0.61	0.76	0.64	0.68	0.60	0.67	0.64
	Recall	0.94	0.94	0.88	0.95	0.93	0.89	0.85	0.89	0.90
	F1	0.76	0.83	0.72	0.84	0.76	0.77	0.70	0.77	0.75
BlueBERT PM	Precision	0.86	0.92	0.92	0.93	0.83	0.89	0.90	0.89	0.88
	Recall	0.93	0.89	0.88	0.93	0.91	0.89	0.86	0.90	0.90
	F1	0.89	0.90	0.90	0.93	0.87	0.89	0.88	0.89	0.89

BERT = Bidirectional Encoder Representations from Transformers; CC = with correction; OD = right eye; OS = left eye; PH = pinhole; SC = without correction.

Discussion

In this study, we identified different types of VA measurements and their lateralities from ophthalmology clinical progress notes, comparing several different pretrained BERT models that were fine-tuned to our task in a weakly supervised manner. Bidirectional Encoder Representations from Transformers models performed with microaveraged F1 scores ranging from 0.75 to 0.90 on the weakly supervised test set and 0.84 to 0.92 on the human-annotated test set. The most common model errors included labeling VA outside of the examination section of the note and missing nonnumeric VA values such as "hand motion" or "light perception."

To our knowledge, this is the first instance of training a deep learning model to recognize VA from free-text clinical notes. Previous studies have sought to identify VA from free-text notes using rule-based algorithms on a much smaller scale. A 2016 study by Mbagwu et al[18] developed a rule-based algorithm to abstract VA and evaluated its performance compared with manual chart review of 100 patient notes, with an exact match rate of approximately 80%.

Table 2. Model Performance on Human-Annotated Test Set

		OD				OS				
		CC		SC		CC		SC		
Model	**Metric**	*No PH*	*PH*	*No PH*	*PH*	*No PH*	*PH*	*No PH*	*PH*	**Microaveraged**
Baseline model	Precision	0.95	0.71	0.94	0.80	1.00	0.95	0.99	1.00	0.93
	Recall	0.81	0.88	0.82	0.94	0.69	0.44	0.75	0.56	0.76
	F1	0.88	0.79	0.88	0.86	0.82	0.60	0.86	0.72	0.83
DistilBERT	Precision	0.74	0.91	0.93	0.99	0.73	1.00	0.96	1.00	0.86
	Recall	0.98	0.75	0.97	0.92	0.98	0.75	0.97	0.88	0.95
	F1	0.84	0.82	0.95	0.95	0.83	0.86	0.96	0.94	0.90
BioBERT	Precision	0.94	0.93	0.79	0.85	0.93	1.00	0.86	0.88	0.87
	Recall	0.87	0.70	0.93	0.78	0.83	0.62	0.75	0.42	0.81
	F1	0.90	0.80	0.85	0.81	0.88	0.77	0.80	0.57	0.84
ClinicalBERT	Precision	0.72	0.97	0.99	0.97	0.70	0.99	0.98	0.98	0.87
	Recall	0.93	0.97	0.93	0.98	0.95	0.99	0.94	0.98	0.95
	F1	0.81	0.97	0.96	0.98	0.81	0.99	0.96	0.98	0.91
BlueBERT PM	Precision	0.73	1.00	0.97	0.98	0.73	1.00	0.97	0.97	0.87
	Recall	0.95	0.97	0.97	1.00	0.97	0.98	0.98	0.98	0.97
	F1	0.83	0.98	0.97	0.99	0.83	0.99	0.97	0.98	0.92

BERT = Bidirectional Encoder Representations from Transformers; CC = with correction; OD = right eye; OS = left eye; PH = pinhole; SC = without correction.

However, this algorithm was developed to work only within the VA section of the EHR note rather than the entire note, limiting its utility. Another study by Baughman et al[19] developed the regular expression-based "Total VA Extraction Algorithm (TOVA)" in Ruby, applying it to ophthalmology consult notes and reporting a 95% concordance between TOVA and manual review of 644 notes. The most common mistake made by the TOVA was identifying VA outside of the examination section, similar to our models. Ophthalmology consult notes may also contain VAs, which are collected under somewhat different or less standardized conditions than in the clinic because consults are usually performed at the bedside in the emergency department or the hospital; thus, they may contain many near VAs rather than distance acuities, and generalizability of this algorithm may be limited. We also developed a rule-based model in our study to compare with the BERT models that performed NER on the human-annotated notes with a microaveraged F1 score of 0.83, precision of 0.93, and recall of 0.76. Whereas these prior studies extracted VA values alone (e.g., "20/50") or VA values with laterality (e.g., "20/30, OS"), our models identify both the VA value and type of VA (e.g., "20/20, OD, without correction, with pinhole"). This allows us to more precisely characterize or filter model outputs before downstream applications. Rule-based algorithms have some limitations which could be overcome by deep learning approaches. For example, rule-based algorithms require advanced domain knowledge to manually encode each possible case, whereas deep learning models learn representations of human language, allowing generalizability to new cases. There is great variation in how VA is documented in EHRs: the type may precede or follow the measurement, sometimes all OD VAs are reported before OS VAs, or OS and OD VAs are reported back-to-back. The high precision of our rule-based model indicated that extracted visual acuities tended to be correct, whereas the low recall was driven by "missed" visual acuities in the text; hand-crafted rules could not capture every possible context in which a VA could be stated.

Our deep learning algorithms outperformed rule-based algorithms in identifying VA from ophthalmology progress notes, with a level of performance similar to previous reports of BioBERT's performance on biomedical NER tasks.[3] In previous studies, BioBERT achieved F1 scores ranging from 72.8 to 93.7 on 9 biomedical NER tasks.[3] Similarly, ClinicalBERT has been shown on 3 i2b2 NER tasks to achieve accuracy ranging from 79.5% to 92.6%.[5] Across 3 NER benchmark data sets, BlueBERT achieved F1 scores ranging from 77.1 to 92.4.[6] In our study, the best-performing model on the task of recognizing different types of VA was BlueBERT, with a microaveraged F1 score of 0.92 on human-annotated notes. This is potentially because this model was trained on both PubMed and the MIMIC-III data set, learning representations for the language of biomedical literature and EHRs. Interestingly, DistilBERT seemed to outperform BioBERT (microaveraged F1 of 0.90 vs. 0.84, respectively), despite only having been pretrained on BooksCorpus and Wikipedia. Unlike the other BERT models in this study, DistilBERT requires fewer computational resources and thus trains more quickly.[7,8] Therefore, it may be beneficial in future studies to pretrain a more compact model like DistilBERT on biomedical corpora similar to BlueBERT.

Unlike rule-based algorithms, deep learning models require large, high-quality training corpora, which may be challenging to access or develop, especially in clinical domains, such as ophthalmology.[20] Accordingly, one unique strength of our study was training on a large corpus of ophthalmology notes using weak supervision to bypass time-consuming domain-specific manual annotation. Our study leveraged semistructured EHR fields to develop a large training corpus for our VA-recognition models, an approach we extended from our previously developed models that identified the slit lamp examination (e.g., conjunctiva, sclera, and cornea) and the fundus examination (e.g., macula and cup-to-disc ratio) from ophthalmology notes.[8] In this present study, the VA-recognition BERT models generally performed well on the human-labeled test set, despite having been trained entirely on algorithmically labeled data. In some cases, VA-recognition BERT models were able to "transcend" the noise in the training labels to perform even better compared with human-annotated ground truth, up to F1 scores of 0.92. It is noteworthy that the VA-recognition baseline regex model was much better at detecting VA than the slit lamp and fundus examination recognition models were at detecting eye examination findings in our previous study, because there is significantly more lexical variability across the 12 components of the slit lamp and fundus examination (e.g., "AC narrow angle" or "Lens 3+ ACC, 2+ NS"), whereas VA is most commonly presented as a numeric fraction (e.g. "20/40").[8] Nonetheless, slit lamp and fundus examination recognition BERT models also outperformed a rule-based model on their tasks, with microaveraged F1 scores ranging from 0.87 (ClinicalBERT) to 0.90 (BioBERT) vs. 0.72 (rule-based model). Thus, even on what seem to be a lexically simpler task (recognizing VA rather than eye examination findings) BERT models outperform hand-crafted rule-based algorithms in entity recognition.

Overall, our study can have implications for facilitating the study of ophthalmic diseases at larger scales by reducing dependence on manual data extraction of VA when such data is sequestered in free-text progress notes. By enabling rapid and automated extraction of VA data from EHR notes, our approach could reduce the need for time-consuming and error-prone human chart review and reduce the need for tedious construction of brittle, rule-based approaches to detect VA. Our models could be used at other EHR systems in which VA is recorded entirely in free text, such as for the Veterans Health Affairs system, which is an area of future research. Our approach would thus mitigate the need for specialized data fields for VA, streamlining the data acquisition processes for downstream ophthalmology research applications.

Our approach has several key limitations related to our training corpus and architecture of the language models. Our data were derived from a single institution, which may limit

the generalizability of our models to EHR notes from other institutions, who may wish to further fine-tune NER models to their unique corpora. Our corpus only contained distance VA (measured via eye chart on a wall 20 feet away), whereas near VA (measured via hand-held card) is also commonly measured in other contexts, such as bedside evaluations. Additionally, our corpus does not contain instances of VA documented from both eyes simultaneously, e.g., "the patient's visual acuity was 20/20 OU," such as might be found in long-form prose letters or dictations. Also, by nature of how the eye examination is performed, some VAs were less represented in our data set, such as pinhole acuities, which are only measured in some circumstances, which could lead to variability in model performance across different VA types. Other limitations pertain to our deep learning models. Because BERT models require inputs to not exceed a specific length (e.g., 128-word pieces for BlueBERT), we split notes into shorter segments to fit model input lengths; however, splitting may cut off the context of the measurements, potentially resulting in reduction in performance for OS visual acuities, typically stated in the text further away from the beginning of the VA section. An alternative approach may involve a long document version of BERT, such as using CogLTX,[21] but this would lose the advantage of pretraining on biomedical corpora and increase the computational resources required. Finally, unlike rule-based algorithms, in which model function is explicitly coded, deep learning models, such as the BERT models, in this study, suffer from diminished interpretability (understanding how the models produced their outputs). To overcome this limitation and examine model outputs, we used the Prodigy tool to evaluate and highlight common errors.

In conclusion, we were able to fine-tune pretrained BERT-based deep learning models to recognize different types of VA from free-text ophthalmology clinical notes using weakly supervised labels. The model based on BlueBERT outperformed those based on BioBERT, ClinicalBERT, DistilBERT, and a rule-based model. Future work is needed to improve the quality of the training labels to further improve the model performance, evaluate model performance on external data sets from multiple institutions, and apply the model to clinical research studies.

Footnotes and Disclosures

Originally received: April 24, 2023.
Final revision: June 20, 2023.
Accepted: July 13, 2023.
Available online: July 19, 2023. Manuscript no. XOPS-D-23-00081R1.

[1] Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, California.

[2] Department of Ophthalmology, Weill Cornell Medicine, New York, New York.

Presented as a Poster at Future Vision Forum in Los Angeles, California on October 31, 2022, and ARVO Annual Meeting in New Orleans, Louisiana on April 27, 2023.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have made the following disclosures: I.B.: Travel expenses — Association for Research in Vision and Ophthalmology (Knights Templar Eye Foundation Travel Grant).

Supported by the National Eye Institute (grant no.: 1K23EY03263501 [to S.Y.W.]); Career Development Award from Research to Prevent Blindness (to S.Y.W.); unrestricted departmental grant from Research to Prevent Blindness (all authors); departmental grant National Eye Institute P30-EY026877 (all authors). The sponsors or funding organizations had no role in the design or conduct of this research.

HUMAN SUBJECTS: No human subjects were included in this study. This study was approved by the Stanford University institutional review board. The institutional review board granted a waiver of informed consent given the scale of the data and observational nature of the data set. The study adhered to the tenets of the declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Bernstein, Koornwinder, Hwang, Wang

Data collection: Bernstein, Koornwinder, Hwang, Wang

Analysis and interpretation: Bernstein, Koornwinder, Hwang, Wang

Obtained funding: Wang

Overall responsibility: Bernstein, Koornwinder, Hwang, Wang.

Abbreviations and Acronyms:
BERT = Bidirectional Encoder Representations from Transformers; **EHR** = electronic health record; **NER** = named entity recognition; **OD** = right eye; **OS** = left eye; **TOVA** = Total VA Extraction Algorithm; **VA** = visual acuity.

Keywords:
Deep learning, Electronic health records, Natural language processing, Ophthalmology, Visual acuity.

Correspondence:
Sophia Y. Wang, 2370 Watson Ct, Palo Alto, CA 94030. E-mail: sywang@stanford.edu.

References

1. Daiber HF, Gnugnoli DM. Visual acuity. In: *StatPearls*. StatPearls Publishing; 2022. http://www.ncbi.nlm.nih.gov/books/NBK563298/. Accessed November 26, 2022.
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pretraining of deep bidirectional transformers for language understanding. Published online May 24, 2019. https://doi.org/10.48550/arXiv.1810.04805
3. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234–1240. https://doi.org/10.1093/bioinformatics/btz682.
4. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. Published online November 28, 2020. https://doi.org/10.48550/arXiv.1904.05342

5. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. Published online June 18, 2019. https://doi.org/10.48550/arXiv.1906.05474
6. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Published online February 29, 2020. https://doi.org/10.48550/arXiv.1910.01108
7. Abadeer M. Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2020:158−167. https://doi.org/10.18653/v1/2020.clinicalnlp-1.18.
8. Wang SY, Huang J, Hwang H, et al. Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam. *Int J Med Inform*. 2022;167:104864. https://doi.org/10.1016/j.ijmedinf.2022.104864.
9. Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: state-of-the-art natural language processing. Published online July 13, 2020. https://doi.org/10.48550/arXiv.1910.03771
10. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE − an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*. 2009;2009:391−395.
11. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. Published online 2009 https://github.com/nltk/nltk. Accessed November 14, 2022.
12. Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. Published online May 23, 1995. https://doi.org/10.48550/arXiv.cmp-lg/9505040
13. Song X, Salcianu A, Song Y, et al. Fast WordPiece tokenization. Published online October 5, 2021. https://doi.org/10.48550/arXiv.2012.15524
14. Prodigy 101 − everything you need to know. Prodigy. https://prodi.gy/docs. Accessed November 26, 2022.
15. Akiba T, Sano S, Yanase T, et al. Optuna: a next-generation hyperparameter optimization framework. Published online July 25, 2019. https://doi.org/10.48550/arXiv.1907.10902
16. Nakayama H. seqeval. Published online November 24, 2022 https://github.com/chakki-works/seqeval. Accessed November 27, 2022.
17. Bernstein IA, Koornwinder A, Wang SY. eyelovedata/ophnotes-ner-va: v1.0.0. Published online March 27, 2023. https://doi.org/10.5281/zenodo.7776114
18. Mbagwu M, French DD, Gill M, et al. Creation of an accurate algorithm to detect Snellen best documented visual acuity from ophthalmology electronic health record notes. *JMIR Med Inform*. 2016;4:e14. https://doi.org/10.2196/medinform.4732.
19. Baughman DM, Su GL, Tsui I, Lee CS, Lee AY. Validation of the total visual acuity extraction algorithm (TOVA) for automated extraction of visual acuity data from free text, unstructured clinical records. *Transl Vis Sci Technol*. 2017;6:2. https://doi.org/10.1167/tvst.6.2.2.
20. Chen JS, Baxter SL. Applications of natural language processing in ophthalmology: present and future. *Front Med (Lausanne)*. 2022;9:906554.
21. Ding M, Zhou C, Yang H, Tang J. CogLTX: applying BERT to long texts. In: *Advances in Neural Information Processing Systems*. Vol 33. Curran Associates, Inc.; 2020:12792−12804. https://proceedings.neurips.cc/paper/2020/hash/96671501524948bc3937b4b30d0e57b9-Abstract.html. Accessed January 20, 2023.