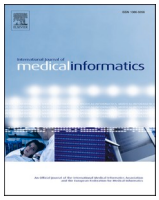
International Journal of Medical Informatics 167 (2022) 104864

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam

Sophia Y. Wang [a,*], Justin Huang [b], Hannah Hwang [c], Wendeng Hu [a], Shiqi Tao [a], Tina Hernandez-Boussard [d]

[a] Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, CA, USA
[b] Johns Hopkins School of Medicine, Baltimore, MD, USA
[c] Department of Ophthalmology, Weill Cornell Medicine, New York, NY, USA
[d] Center for Biomedical Informatics Research, Stanford University, Palo Alto, CA, USA

ARTICLE INFO

Keywords:
Natural language processing
Named entity recognition
Weak supervision
Deep learning
Ophthalmology
Electronic health records

ABSTRACT

Objective: To develop deep learning models to recognize ophthalmic examination components from clinical notes in electronic health records (EHR) using a weak supervision approach.
Methods: A corpus of 39,099 ophthalmology notes weakly labeled for 24 examination entities was assembled from the EHR of one academic center. Four pre-trained transformer-based language models (DistilBert, BioBert, BlueBert, and ClinicalBert) were fine-tuned to this named entity recognition task and compared to a baseline regular expression model. Models were evaluated on the weakly labeled test dataset, a human-labeled sample of that set, and a human-labeled independent dataset.
Results: On the weakly labeled test set, all transformer-based models had recall > 0.93, with precision varying from 0.815 to 0.843. The baseline model had lower recall (0.769) and precision (0.682). On the human-annotated sample, the baseline model had high recall (0.962, 95 % CI 0.955–0.067) with variable precision across entities (0.081–0.999). Bert models had recall ranging from 0.771 to 0.831, and precision >=0.973. On the independent dataset, precision was 0.926 and recall 0.458 for BlueBert. The baseline model had better recall (0.708, 95 % CI 0.674–0.738) but worse precision (0.399, 95 % CI -0.352–0.451).
Conclusion: We developed the first deep learning system to recognize eye examination components from clinical notes, leveraging a novel opportunity for weak supervision. Transformer-based models had high precision on human-annotated labels, whereas the baseline model had poor precision but higher recall. This system may be used to improve cohort and feature identification using free-text notes.Our weakly supervised approach may help amass large datasets of domain-specific entities from EHRs in many fields.

1. Introduction

More information than ever is stored in free-text notes within the electronic health record (EHR), including descriptions of patients' symptoms, examination, assessment, and plan. However, annotating this data and identifying clinical cohorts with shared characteristics for research are time-consuming and expensive tasks that require domain expertise [1,2]. Automated annotation and extraction of this information could provide the basis for efficiently defining research cohorts by phenotype or treatment trajectory and developing predictive models for patient outcomes. Especially critical in ophthalmology is the eye examination, which encompasses patients' anterior and posterior eye segment exams, including features not commonly documented in structured billing codes, such as disc hemorrhages for glaucoma patients, geographic atrophy for macular degeneration patients, or the signs of previous eye surgery or laser. Without fast methods to collect

Abbreviations: EHR, electronic health record; NLP, natural language processing; BERT, Bidirectional Encoder Representations from Transformers; NER, named entity recognition; SLE, slit lamp exam; FE, fundus exam; Pr, precision; Rec, recall; L/L, lids and lashes; C/S, conjunctiva and sclera; K, cornea; AC, anterior chamber; AV, anterior vitreous; Disc, optic disc; CDR, cup to disc ratio; M, macula; V, vessels; P, periphery.

* Corresponding author.
E-mail address: sywang@stanford.edu (S.Y. Wang).

https://doi.org/10.1016/j.ijmedinf.2022.104864
Received 16 May 2022; Received in revised form 11 August 2022; Accepted 5 September 2022
Available online 16 September 2022

these critical findings from free-text notes, the abilities to characterize ophthalmology cohorts and build predictive models are severely hampered.

A major challenge is the unstructured nature of clinical free-text notes which require natural language processing (NLP) techniques to understand and compute over. Transformer-based deep learning architectures such as BERT (Bidirectional Encoder Representations from Transformers) [3] and its descendants have ushered in a performance revolution on many NLP tasks, including named entity recognition (NER). Versions of BERT pretrained on biomedical literature, such as ClinicalBert [4,5], BlueBert [6], and BioBert [7],. as well as reduced versions like DistilBert [8], have performed well on biomedical named entity tasks such as recognizing disease entities [4–7,9]. However, there have not been efforts to use these NLP models to perform NER for ophthalmology exam components. The most advanced previous work extracted visual acuity measurements from free-text progress notes using a rule-based text-processing approach [10]. Thus, we tested these 4 models for our ophthalmology NER task.

A critical barrier to the use of BERT-based models to perform NER in ophthalmology is the lack of large annotated corpora for model training, a barrier shared by everyone desiring to use these models for real-world or domain-specific tasks outside of benchmark datasets. Manual annotations on clinical notes require expertise and are time-consuming and difficult to produce on a scale sufficiently large to train BERT models. Thus, this study sought to use a weakly supervised approach requiring minimal manual annotation of training corpora to build and evaluate transformer-based deep learning models that can identify elements of the anterior segment (slit lamp) exam (SLE) and posterior segment (fundus) exam (FE) and their lateralities from ophthalmology clinical notes. The goal was to train models for our domain-specific task using readily available weakly labeled data from EHRs and compare model performance against a smaller subset of human-annotated data. We aimed to develop the first models that researchers could eventually use to characterize ophthalmology patients based on granular clinical findings, using a weak supervision approach that researchers in other specialties could also use to build their own specialized NER systems.

2. Materials and methods

2.1. Data source

We identified from the Stanford Research Repository [11] all encounters with associated progress notes, slit lamp examinations, and fundus examinations of patients seen by the Department of Ophthalmology at Stanford University since 2008. This study was approved by the Stanford University Institutional Review Board.

2.2. Preprocessing labels

SmartForms are used to document semi-structured text in many deployments of this EHR system (Epic), for multiple specialties including ophthalmology, where they are used to document the eye examination. Information in SmartForms can be imported into the note using providers' custom note templates. Thus, pairs of SmartForms and corresponding clinical notes represent notes which are labeled with the clinical exam information (Fig. 1). A summary of the many eye exam components and their abbreviations is given in Supplemental Table 1.

A custom preprocessing pipeline was developed to assign token-level entity labels for each document ("training labels"), illustrated in Fig. 2. Each document was pre-tokenized using the Treebank Tokenizer in the Python Natural Language ToolKit v3.5 [12]. For each component of an eye exam ("label", e.g. "conjunctiva/sclera"), we identified the finding (e.g. "white and quiet") documented in the SmartForm, then searched the patient's note for these tokens and labeled them accordingly. If the search yielded multiple matches (e.g., multiple findings are "normal"), a greedy process was used to assign each label to a token, iterating through each label and assigning the first matched token to that label if the token wasn't already assigned to another label. Labels were constructed in the Inside-Outside-Beginning (IOB2) format [13,14], with 'O' for no label or outside of the entity, 'B-label' for the beginning of an entity, and 'I-label' for tokens that continue (or are inside of) an entity. The result of this process is a list of tokens and a list of corresponding SLE or FE labels. Data were split into 80:10:10 train, validation, and test sets (N = 31279, 3910, 3910 respectively). Full notes were split into shorter subdocuments for input into models with maximum input lengths. BERT

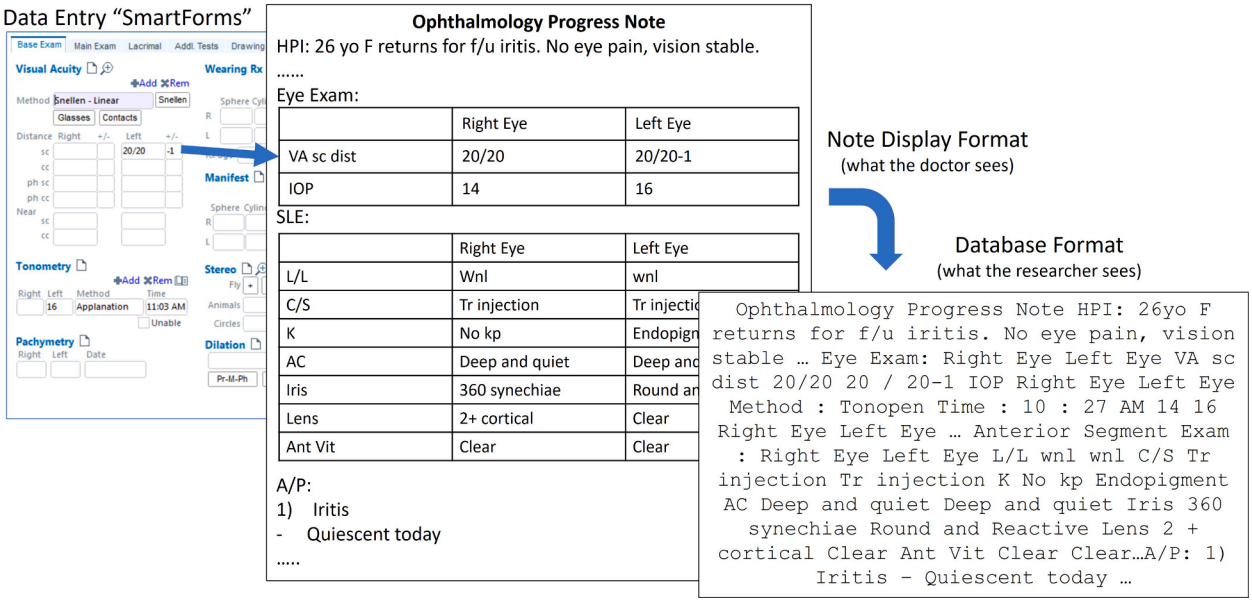


Fig. 1. Example SmartForm and corresponding clinical progress note Legend: The leftmost panel shows the SmartForm template which clinicians use to enter text documenting different parts of the eye exam into discrete labeled fields. This information can then be imported via customizable templates into each clinician's progress notes. The progress notes are then stored into a research database. VA = visual acuity; sc = sans correction; IOP = intra-ocular pressure; L/L = lids and lashes; C/S = conjunctiva and sclera; K = cornea; AC = anterior chamber; Ant Vit = anterior vitreous; HPI = history of present illness; f/u = follow-up.

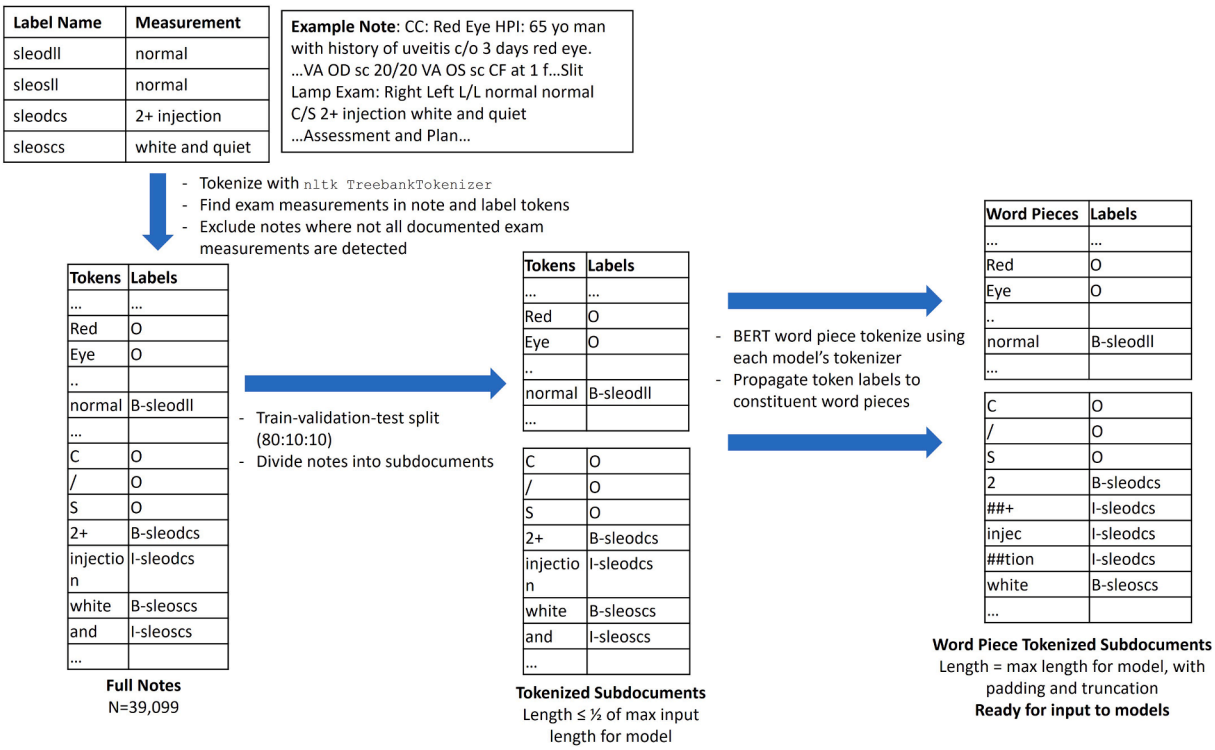


Fig. 2. Preprocessing pipeline for clinical progress notes and corresponding SmartForm entity labels Legend: An example progress note and its corresponding SmartForm documentation is shown, as well as the process by which SmartForm labeled entities are assigned to individual words in the progress note. Notes with individual words labeled as entities are tokenized, split into shorter subdocuments, and word piece tokenized as appropriate for input into each Bert model. Label Name: entity label describing a portion of the eye examination. Measurement: The measurement associated with an examination component. Token: A single element of text for computational processing. Labels: entity labels assigned to each token for computational processing. sleodll = slit lamp exam, right eye, lids and lashes; sleosll = slit lamp exam, left eye, lids and lashes; sleodcs = slit lamp exam, right eye, conjunctiva and sclera; sleoscs = slit lamp exam, left eye, conjunctiva and sclera.

word piece tokenization was performed and original full word token-level labels were propagated to each word piece token as appropriate. Padding and truncation were used to standardize the length of each subdocument for input into models.

2.3. Baseline classifier

The baseline NER model uses regular expressions to search for anatomical "header" keywords within the text and assigns the text sandwiched between headers to the eye exam component indicated by the preceding header. Supplemental Table 2 summarizes the overall approach and specific regular expressions.

2.4. BERT modeling and experimental details

BERT models trained in this study include biomedical extensions of BERT: BioBert, pretrained on all PubMed abstracts [7], ClinicalBert, pretrained on PubMed abstracts, PubMed central, and MIMIC critical care notes [5,6], and BlueBert pretrained on PubMed abstracts and MIMIC notes for more steps than ClinicalBert [6]. We also tested DistilBert, a smaller version of BERT with 40 percent fewer parameters shown to perform well on NER tasks, to investigate whether a light-weight model pre-trained on general English language corpora could perform comparably [8].

All pre-trained models were initialized through the huggingface transformers library [15] for the token classification task and fine-tuned to identify the eye examination entities. All models were trained with standard cross-entropy loss function for token classification, with the Adam optimizer on NVIDIA Tesla P100 GPU, and with the maximum allowable input lengths for each model. Validation loss was calculated after each epoch, and early stopping was used with patience of 3. The model with the best validation loss was used for final evaluation, with final tuned best training parameters summarized in Supplemental Table 3.

2.5. Standard evaluation metrics

We evaluated performance using the Python `seqeval` package (v1.2.2)[16] using standard metrics for NER: precision, recall, and F1 score for each named entity type as well as micro-averaged metrics across all entities. These metrics are commonly reported for many biomedical NER tasks and allow the reader to judge which approaches may be most suitable for their applications [17,18]. Confidence intervals (95 % CI) were obtained via bootstrapping with 1000 replicates.

2.6. Error analysis and qualitative evaluation metrics

We analyzed the performance of each model on a sample of 300 manually annotated documents from the test set, using the Prodigy annotation tool (https://prodi.gy/) to visualize and correct our models' predictions. We also manually corrected and analyzed a sample of the models' training labels (N = 100), given that these were assigned algorithmically and could contain noise. On the annotated sample of documents, we report standard evaluation metrics for each models' predictions compared to the ground truth human annotation. We identified qualitative examples of typical contexts in which the models fail.

We also identified an independent set of clinical progress notes ("outset") in which providers directly typed their eye exam findings in free text rather than into SmartForms. The format of these free text exam findings was more variable and more customized to individual providers. Of the 84,292 notes in the "outset", we randomly sampled 300 to investigate in detail. We divided these notes into shorter subdocuments

and input them into the BlueBert model, then manually reviewed the model predictions in Prodigy to identify model strengths and weaknesses on an independent set of fully free text notes. Manual annotation was carried out by a clinician with > 8 years of experience in the ophthalmic field.

2.7. Model demonstration and code release

We have built a demonstration for all models which is available online (https://flasknerapp.herokuapp.com/). Code for developing these models has been publicly released [19].

3. Results

3.1. Model performance against training labels

Performance of the models against the weakly supervised SmartForm-based training labels in the test set is summarized in Table 1 with 95 % confidence intervals in Supplemental Table 4. All BERT models had excellent recall, micro-averaging over 0.93. Precision varied, ranging from a high of 0.843 micro-averaged for BioBert to a low of 0.815 micro-averaged for ClinicalBert. The baseline model had lower recall (0.769) and precision (0.682) than the BERT models.

3.2. Common sources of training label errors

During the human annotation of the training labels, common sources of errors were noted in the weak label-assignment. Performance metrics of the labeling algorithm compared to human-annotated ground truth is presented in Supplemental Table 5. The overall micro-averaged F1 score of training labels when evaluated against human annotations was high (0.957 [95 % CI 0.944–0.969]). Examples of common types of errors are illustrated in Supplemental Fig. 1, including instances where some exam components are not labeled or partially missed and where text outside of the eye exam is mislabeled as an eye exam finding.

3.3. Model performance against human-annotated ground truth

Given the noise in the training labels, models' predictions were compared against notes in the test set human-annotated for ground truth to evaluate the true performance (Table 2, confidence intervals in Supplemental Table 6). The baseline model had high micro-averaged recall (0.962 [95 %CI 0.955–0.067]) but precision was variable, ranging from 0.081 to 0.999 depending on specific entity, for an overall micro-averaged precision of 0.720 [95 %CI 0.704–0.737]. BERT models had uniformly better performance, with micro-averaged F1-score ranging from 0.863 [95 % CI 0.853–0.878] (BioBert) to 0.903 [95 % CI 0.895–0.912] (BlueBert), micro-averaged recall ranging from 0.771 [95 %CI 0.756–0.785] (BioBert) to 0.831 [95 %CI 0.818–0.843] (BlueBert), and micro-averaged precision ranging from 0.973 [95 % CI 0.963–0.982] (ClinicalBert) to 0.990 [95 %CI 0.985–994] (BlueBert). Thus, for BERT models, precision was higher than recall when evaluated against human-annotated ground truth data, while recall was higher than precision when tested on algorithmically-labeled data. Most model errors occur when contiguous words that comprise a finding are not all labeled, with some words "skipped" in labeling (Fig. 3).

3.4. Evaluation against independent fully free-text clinical notes

To evaluate whether a model trained on SmartForm-labeled notes could generalize to completely free-text progress notes, we evaluated the baseline model and BlueBert, which had the best micro-averaged F1 score on the human-annotated test set, on an independent set of notes which document the eye examination in free-text. This approach is possible at our center because many ophthalmologists prefer not to use SmartForm-driven templates to document their findings; thus, their notes are examples of completely free-text ophthalmology notes. Broadly, the BlueBert model had very good precision (micro-average 0.926) but relatively poor recall (0.458), whereas the baseline model had better recall (0.708) but poor precision (0.399) (Table 3). Many common model errors (Fig. 4) could be attributed to the significantly

Table 1
Model Performance Against Weakly Supervised Labels in Test Set.

Component		Baseline Model			DistilBert			BioBert			BlueBert			ClinicalBert		
		Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1
SLE, Right	L/L	0.943	0.905	0.924	0.967	0.983	0.975	0.979	0.988	0.983	0.974	0.983	0.979	0.960	0.969	0.964
	C/S	0.871	0.908	0.889	0.877	0.972	0.922	0.900	0.983	0.940	0.907	0.988	0.946	0.899	0.989	0.942
	K	0.841	0.879	0.860	0.860	0.966	0.910	0.891	0.988	0.937	0.877	0.981	0.926	0.872	0.969	0.918
	AC	0.882	0.935	0.908	0.893	0.970	0.930	0.920	0.994	0.955	0.918	0.993	0.954	0.906	0.988	0.945
	Iris	0.876	0.927	0.901	0.906	0.979	0.941	0.921	0.993	0.956	0.917	0.991	0.952	0.914	0.991	0.951
	Lens	0.724	0.831	0.774	0.804	0.970	0.880	0.824	0.985	0.897	0.810	0.977	0.886	0.763	0.935	0.840
	AV	0.442	0.583	0.503	0.846	0.971	0.904	0.855	0.980	0.913	0.749	0.862	0.801	0.837	0.967	0.897
FE, Right	Disc	0.689	0.775	0.730	0.774	0.903	0.834	0.804	0.953	0.872	0.780	0.948	0.856	0.739	0.922	0.821
	CDR	0.411	0.516	0.457	0.785	0.903	0.840	0.773	0.969	0.860	0.770	0.929	0.842	0.773	0.928	0.844
	M	0.574	0.677	0.621	0.725	0.888	0.798	0.741	0.907	0.816	0.719	0.878	0.790	0.707	0.897	0.790
	V	0.715	0.760	0.737	0.795	0.886	0.838	0.795	0.889	0.839	0.780	0.866	0.821	0.754	0.834	0.792
	P	0.483	0.582	0.528	0.678	0.834	0.748	0.702	0.860	0.773	0.674	0.806	0.734	0.688	0.819	0.748
SLE, Left	L/L	0.931	0.914	0.922	0.952	0.983	0.967	0.963	0.991	0.977	0.954	0.984	0.969	0.956	0.983	0.969
	C/S	0.866	0.910	0.887	0.872	0.966	0.916	0.901	0.986	0.942	0.891	0.979	0.933	0.861	0.970	0.912
	K	0.817	0.878	0.847	0.850	0.955	0.899	0.876	0.981	0.926	0.870	0.975	0.920	0.847	0.955	0.898
	AC	0.869	0.933	0.900	0.898	0.981	0.937	0.911	0.989	0.948	0.902	0.985	0.942	0.904	0.986	0.944
	Iris	0.871	0.929	0.899	0.909	0.985	0.945	0.920	0.994	0.956	0.915	0.988	0.950	0.904	0.984	0.942
	Lens	0.714	0.830	0.767	0.793	0.954	0.866	0.812	0.979	0.888	0.797	0.964	0.873	0.781	0.955	0.859
	AV	0.008	0.011	0.009	0.845	0.959	0.898	0.849	0.973	0.907	0.778	0.903	0.836	0.798	0.922	0.855
FE, Left	Disc	0.673	0.774	0.720	0.742	0.839	0.788	0.779	0.917	0.842	0.736	0.866	0.796	0.731	0.921	0.815
	CDR	0.408	0.518	0.457	0.782	0.884	0.830	0.765	0.974	0.857	0.771	0.912	0.835	0.766	0.899	0.827
	M	0.568	0.671	0.615	0.711	0.825	0.763	0.732	0.857	0.790	0.708	0.844	0.770	0.657	0.844	0.739
	V	0.704	0.766	0.734	0.804	0.861	0.831	0.795	0.895	0.842	0.812	0.839	0.826	0.765	0.849	0.805
	P	0.338	0.478	0.396	0.677	0.776	0.723	0.697	0.849	0.766	0.665	0.782	0.719	0.657	0.800	0.722
Micro-avg		0.682	0.769	0.723	0.828	0.931	0.876	**0.843**	**0.957**	**0.896**	0.827	0.933	0.877	0.815	0.934	0.870

Pr = precision, Rec = recall, SLE = slit lamp exam, FE = fundus exam, L/L = lids and lashes, C/S = conjunctiva and sclera, K = cornea, AC = anterior chamber, AV = anterior vitreous, Disc = optic disc, CDR = cup to disc ratio, M = macula, V = vessels, P = periphery.

Table 2
Model Performance Against Human-Annotated Ground Truth in Test Set.

Component		Baseline Model			DistilBert			BioBert			BlueBert			ClinicalBert		
		Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1
SLE, Right	**L/L**	0.988	0.985	0.987	0.997	0.771	0.870	0.886	0.759	0.818	0.974	0.775	0.863	0.986	0.783	0.873
	C/S	0.991	0.987	0.989	1.000	0.926	0.962	0.991	0.892	0.939	0.995	0.937	0.965	0.980	0.935	0.957
	K	0.964	0.978	0.971	0.995	0.769	0.867	0.967	0.768	0.857	0.970	0.785	0.868	0.979	0.736	0.840
	AC	0.999	0.991	0.995	1.000	0.970	0.985	1.000	0.943	0.971	0.997	0.975	0.986	0.988	0.979	0.984
	Iris	0.995	0.993	0.994	1.000	0.923	0.960	0.978	0.863	0.917	0.995	0.931	0.962	0.988	0.930	0.959
	Lens	0.977	0.985	0.981	0.998	0.658	0.793	0.992	0.651	0.786	0.994	0.738	0.847	0.988	0.708	0.825
	AV	0.662	0.918	0.769	1.000	0.639	0.779	1.000	0.545	0.706	1.000	0.600	0.750	0.977	0.632	0.767
FE, Right	**Disc**	0.988	0.968	0.978	0.968	0.733	0.834	0.961	0.689	0.802	0.963	0.802	0.875	0.934	0.786	0.854
	CDR	0.583	0.661	0.620	0.943	0.878	0.909	0.980	0.882	0.928	1.000	0.965	0.982	1.000	0.953	0.976
	M	0.945	0.961	0.953	0.987	0.643	0.779	0.984	0.547	0.703	1.000	0.653	0.790	0.977	0.706	0.820
	V	0.991	0.952	0.971	0.934	0.855	0.893	0.937	0.722	0.816	0.892	0.823	0.856	0.929	0.862	0.894
	P	0.973	0.827	0.894	0.977	0.768	0.860	0.980	0.698	0.815	1.000	0.741	0.851	0.921	0.767	0.837
SLE, Right	**L/L**	0.989	0.987	0.988	1.000	0.785	0.887	0.967	0.779	0.863	1.000	0.794	0.885	0.983	0.796	0.880
	C/S	0.986	0.991	0.988	1.000	0.946	0.887	1.000	0.915	0.956	0.994	0.950	0.972	0.966	0.955	0.960
	K	0.960	0.990	0.975	1.000	0.791	0.887	0.974	0.790	0.872	1.000	0.783	0.879	0.987	0.751	0.844
	AC	0.986	0.995	0.991	1.000	0.979	0.989	1.000	0.942	0.970	1.000	0.982	0.991	0.995	0.969	0.982
	Iris	0.985	0.995	0.990	1.000	0.937	0.887	0.987	0.900	0.941	1.000	0.931	0.964	0.985	0.938	0.844
	Lens	0.953	0.988	0.970	1.000	0.667	0.887	1.000	0.659	0.794	1.000	0.750	0.857	1.000	0.731	0.844
	AV	0.081	0.854	0.149	1.000	0.639	0.887	0.992	0.500	0.665	1.000	0.624	0.769	0.990	0.580	0.732
FE, Right	**Disc**	0.959	0.972	0.966	0.970	0.692	0.808	0.947	0.666	0.782	0.954	0.767	0.851	0.945	0.815	0.875
	CDR	0.711	0.990	0.828	0.980	0.875	0.925	0.988	0.890	0.936	1.000	0.954	0.976	1.000	0.952	0.975
	M	0.905	0.968	0.936	0.992	0.622	0.765	0.970	0.504	0.663	0.992	0.626	0.768	0.938	0.682	0.790
	V	0.963	0.980	0.971	0.949	0.802	0.869	0.953	0.680	0.794	0.944	0.823	0.879	0.956	0.879	0.916
	P	0.508	0.729	0.599	0.965	0.705	0.814	0.973	0.675	0.797	1.000	0.747	0.855	0.909	0.787	0.843
Micro-avg		0.720	0.962	0.824	0.989	0.804	0.887	0.979	0.771	0.863	**0.990**	**0.831**	**0.903**	0.973	0.829	0.895

Pr = precision, Rec = recall, SLE = slit lamp exam, FE = fundus exam, L/L = lids and lashes, C/S = conjunctiva and sclera, K = cornea, AC = anterior chamber, AV = anterior vitreous, Disc = optic disc, CDR = cup to disc ratio, M = macula, V = vessels, P = periphery.

different format of the notes generated in this manner, such as when the exam findings are presented in their entirety-one eye at a time rather than alternating right and left for each anatomical portion.

4. Discussion

We leveraged EHR data captured from routine ophthalmic care to train deep learning NER models in a semi-supervised manner to detect with high precision findings from the anterior and posterior segment exams and their lateralities embedded within ophthalmology clinical progress notes. To our knowledge, this is the first attempt to train a deep learning model to recognize eye examination components from clinical progress notes.

The level of performance we observed for recognition of ophthalmic exam components is within the expected range based on previous reports of BioBert's performance on biomedical NER tasks [7], and all BERT models outperformed the baseline regex model. Hand-crafted regular expressions can be an effective and computationally inexpensive means of recognizing text sequences; however, because note templates vary greatly between doctors, it is infeasible to build regular expressions that capture all patterns. The baseline model's approach to detecting entities is inflexible and only works reliably for notes that document exam findings in a consistent order with all heading keywords present. The baseline model also struggled to distinguish right from left eye findings when these findings were contiguous without obvious syntactical signals indicating the end of the right and the beginning of the left eye finding. In contrast, transformer models can learn to label entities more flexibly. Although the original training labels were noisy, results suggested that BERT models trained using weak labels can ignore some noise to achieve better performance against the human-annotated labels. The baseline model also frequently failed when assigning entities to the final heading of any exam section, as there is no reliable syntactical way to identify the end of the last finding. This limitation exists even when note-taking styles are consistent, thus rendering transformer models favorable to the baseline model in multiple ways.

A unique aspect of our study is evaluating performance on an independent set of notes not generated in the same manner as the original corpus, simulating how ophthalmic notes might appear at other sites or systems. With these notes that were structurally quite dissimilar to notes from the training set, our model maintained excellent precision in many categories. Recall was lower, indicating that the model misses many entities, potentially because they may occur in contexts different from in the training set. The baseline model had generally preserved recall and worse precision, consistent with our findings that the baseline model often over-identifies words as belonging to the eye exam, which may happen when the patterns for the context cues (headers) are missing or different and the regex is allowed to match long sequences of irrelevant text. Both types of models relied on the structure of the presented findings and thus struggled when they were not presented in an alternating right and left fashion. Further work to improve performance may attempt to combine both types of models to optimize both precision and recall, or to gather larger training datasets from a greater variety of institutions and providers.

The ability to start with a relatively large training corpus without manual annotation is the main strength of weak supervision. Rather than manually annotating 24 different eye examination entities across thousands of notes, we algorithmically produced nearly 40,000 relatively cleanly labeled notes. Although crowdsourcing can be a way to generate large annotated corpora for training purposes [20–23], clinical text is difficult to reliably de-identify to the point where release to crowdsourcing workers would be safe. Medical text is also full of abbreviations and would require significant training of crowdsourced workers to interpret.

Our work performing NER for the ophthalmic exam has broad potential applications. Researchers in other specialties may wish to utilize a similar weakly supervised NER approach, for example to automatically recognize polyps from colonoscopy reports, tumors from pathology reports, etc. Reconstruction of tabular structure from unstructured text is another general challenge for which this approach may be used. The granular characterization of patients' clinical characteristics can enable

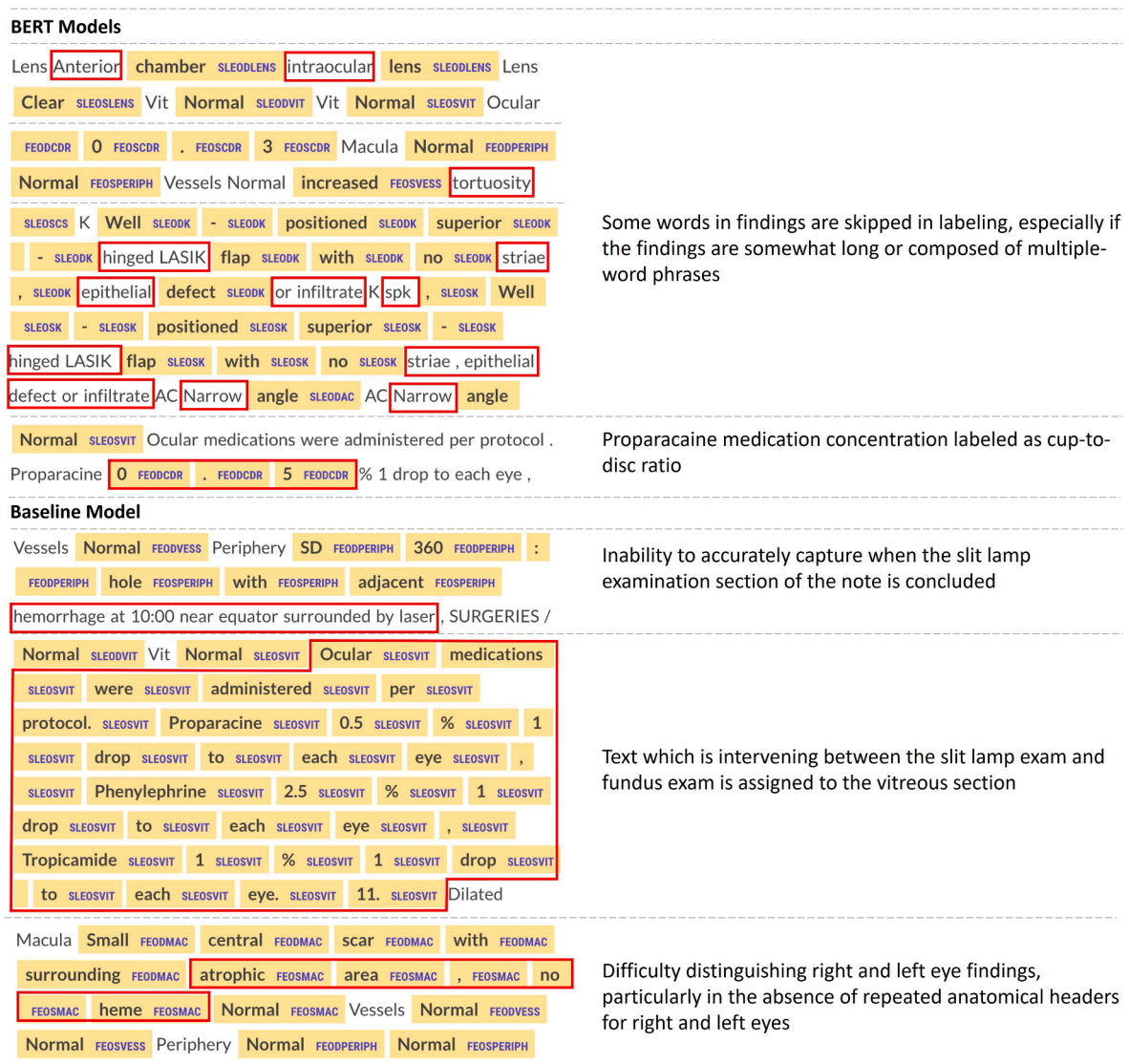


Fig. 3. Examples of Common Types of BERT and Baseline Model Prediction Errors on Test Notes Legend: Examples of common model mistakes on the text from the test set are given in the left column, along with corresponding explanations to the right. Areas highlighted in yellow show the model prediction. Areas boxed in red are those areas where the model has made a mistake, either in the prediction label or in failure to recognize any entity.

cohort construction for observational research and for clinical trials, which often have complex inclusion and exclusion criteria [23]. Similarly, researchers building predictive models for ophthalmic or other clinical outcomes may wish to use NER to construct meaningful input features for predictive models to improve performance and explainability. Many EHR systems are entirely free-text, such as in the Veterans' Health Administration. To understand eye examination components from such systems on a large scale, our NER model can be fine-tuned to suit those ophthalmology notes, and it may require manually annotating fewer notes than training de novo without weak supervision.

Several challenges and limitations remain. The loss of tabular formatting in the notes makes it difficult even for human graders to understand the exam findings, so perhaps it is unsurprising that the models may find it difficult. Great variation exists in how eye exam findings are documented: for example, sometimes with all right eye findings reported and then all left eye findings, and sometimes alternating right and left. These models are also not designed to handle cases of bilateral findings, such as "normal OU." Furthermore, notes must be split up into shorter segments because of model input length limitations. Because the sections documenting eye examination components are not true sentences, notes could be split in awkward locations, cutting off the context of the findings. Finally, the models trained on weakly labeled data sometimes replicated noise present in training labels.

5. Conclusion

In conclusion, we have developed, to our knowledge, the first deep learning pipeline to recognize eye examination components from progress notes. Our system leverages a weakly supervised system to produce nearly 40,000 relatively cleanly labeled notes to train BERT-based models. The models performed better on a set of manually labeled notes than on the algorithmically labeled notes, suggesting the BERT-based models were able to learn entity recognition patterns beyond the small noise present in the training set. Our work holds many potential research applications, from precise cohort design to feature engineering for predictive model development. This weak supervision approach leveraging routinely collected EHR data may also be generalized to other specialties seeking to train models to recognize specific entities from free-text notes.

Table 3
Model Performance on Independent Set of Free-Text Progress Notes.

Component		Baseline Model									BlueBert Model								
		Pr	95 % CI		Rec	95 % CI		F1	95 % CI		Pr	95 % CI		Rec	95 % CI		F1	95 % CI	
SLE, Right	L/L	0.618	0.390	0.890	0.797	0.753	0.832	0.696	0.523	0.851	0.992	0.978	1.000	0.447	0.394	0.489	0.616	0.564	0.655
	C/S	0.838	0.795	0.875	0.803	0.767	0.835	0.820	0.788	0.849	0.964	0.932	0.988	0.526	0.472	0.579	0.681	0.631	0.723
	K	0.756	0.693	0.814	0.830	0.799	0.855	0.791	0.747	0.829	0.840	0.763	0.907	0.466	0.418	0.503	0.599	0.553	0.634
	AC	0.810	0.765	0.848	0.769	0.736	0.799	0.789	0.754	0.818	0.949	0.921	0.975	0.670	0.621	0.715	0.785	0.749	0.817
	Iris	0.770	0.720	0.820	0.764	0.720	0.801	0.767	0.726	0.805	0.930	0.894	0.965	0.605	0.542	0.660	0.733	0.686	0.773
	Lens	0.229	0.188	0.289	0.789	0.750	0.825	0.355	0.303	0.422	0.871	0.807	0.934	0.493	0.447	0.534	0.630	0.583	0.668
	AV	0.014	0.006	0.024	0.166	0.083	0.254	0.025	0.011	0.044	0.500	0.100	1.000	0.032	0.005	0.064	0.060	0.009	0.118
FE, Right	Disc	0.673	0.605	0.736	0.667	0.625	0.708	0.670	0.621	0.713	0.974	0.932	1.000	0.278	0.225	0.332	0.433	0.365	0.495
	CDR	0.438	0.367	0.519	0.498	0.426	0.566	0.466	0.408	0.522	1.000	0.000	1.000	0.009	0.000	0.028	0.018	0.000	0.055
	M	0.866	0.783	0.931	0.690	0.639	0.739	0.768	0.719	0.811	0.943	0.892	0.988	0.331	0.275	0.387	0.490	0.426	0.550
	V	0.900	0.835	0.950	0.697	0.651	0.740	0.785	0.741	0.823	0.984	0.963	1.000	0.540	0.461	0.614	0.697	0.628	0.756
	P	0.770	0.716	0.821	0.592	0.560	0.624	0.669	0.636	0.699	0.926	0.868	0.972	0.545	0.481	0.601	0.686	0.629	0.735
SLE, Left	L/L	0.360	0.174	0.745	0.798	0.750	0.837	0.496	0.284	0.778	0.988	0.975	1.000	0.451	0.404	0.493	0.620	0.573	0.658
	C/S	0.694	0.636	0.751	0.787	0.748	0.823	0.738	0.690	0.779	0.981	0.964	0.996	0.498	0.442	0.547	0.661	0.608	0.705
	K	0.500	0.422	0.574	0.797	0.748	0.836	0.615	0.544	0.678	0.863	0.797	0.922	0.339	0.288	0.382	0.486	0.429	0.535
	AC	0.591	0.528	0.650	0.732	0.687	0.775	0.654	0.602	0.704	0.950	0.921	0.978	0.663	0.618	0.708	0.781	0.746	0.814
	Iris	0.604	0.534	0.672	0.746	0.693	0.789	0.668	0.607	0.721	0.945	0.908	0.981	0.437	0.373	0.497	0.598	0.534	0.650
	Lens	0.241	0.143	0.451	0.765	0.722	0.807	0.366	0.241	0.569	0.903	0.862	0.941	0.500	0.453	0.543	0.644	0.599	0.681
	AV	0.005	0.002	0.008	0.049	0.021	0.085	0.009	0.004	0.015	0.250	0.000	1.000	0.010	0.000	0.025	0.019	0.000	0.048
FE, Left	Disc	0.248	0.195	0.312	0.515	0.438	0.587	0.335	0.272	0.403	0.831	0.701	0.950	0.197	0.152	0.248	0.318	0.255	0.382
	CDR	0.040	0.018	0.066	0.085	0.041	0.136	0.054	0.025	0.088	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	M	0.385	0.282	0.497	0.553	0.445	0.650	0.454	0.351	0.558	0.964	0.922	0.993	0.371	0.309	0.432	0.535	0.467	0.600
	V	0.557	0.469	0.645	0.623	0.546	0.691	0.588	0.512	0.664	0.955	0.916	0.984	0.562	0.488	0.632	0.708	0.644	0.762
	P	0.098	0.071	0.127	0.176	0.130	0.227	0.126	0.092	0.161	0.877	0.819	0.934	0.560	0.516	0.602	0.684	0.639	0.721
Micro-avg		0.399	0.352	0.451	**0.708**	0.674	0.738	0.511	0.466	0.557	**0.926**	0.910	0.941	0.458	0.430	0.485	**0.613**	0.586	0.636

Pr = precision, Rec = recall, 95 % CI = 95 % confidence interval, SLE = slit lamp exam, FE = fundus exam, L/L = lids and lashes, C/S = conjunctiva and sclera, K = cornea, AC = anterior chamber, AV = anterior vitreous, Disc = optic disc, CDR = cup to disc ratio, M = macula, V = vessels, P = periphery.

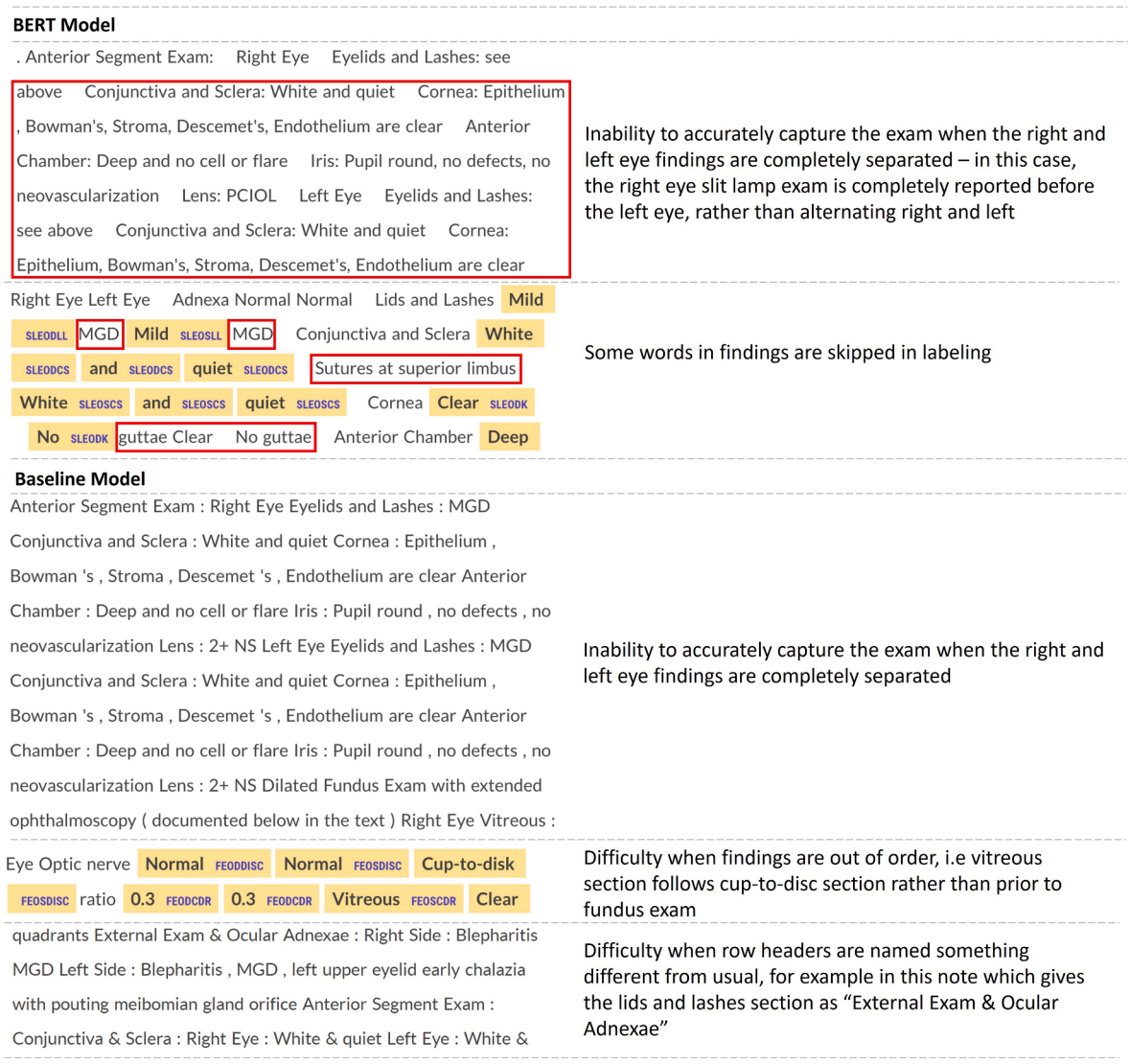


Fig. 4. Examples of Common Types of BERT and Baseline Model Prediction Errors on Independent Notes Legend: Examples of common model mistakes on the text from the independent notes are given in the left column, along with corresponding explanations to the right. Areas highlighted in yellow show the model prediction. Areas boxed in red are those areas where the model has made a mistake, either in the prediction label or in failure to recognize any entity.

Summary table

What is known?	What does this add?
– Many important clinical characteristics of patients are sequestered in unstructured clinical free-text progress notes. – In the field of ophthalmology, there has been little previous work directed towards extracting ophthalmology examination components from free-text progress notes, despite the importance of these findings for cohort identification and characterization. – Bidirectional Encoder Representations from Transformers (BERT) models have enabled a leap in performance in biomedical named entity recognition tasks. – However, training BERT models requires annotated corpora which are difficult to produce on a sufficiently large scale.	– We develop the first deep learning models to identify findings from the ophthalmology exam documented in unstructured ophthalmology progress notes in electronic health records. – We leverage routinely captured data from the electronic health records to develop a weakly supervised approach that amasses a large training corpus with minimal noise and without laborious manual annotation. – Our BERT-based models outperformed a baseline regular-expression based model and also performed better on a manually annotated “ground truth” test set than against the weakly supervised labels.

Authors’ contributions

All authors have made significant contributions to the manuscript. SYW and JH designed the study, analyzed and extracted the data, and performed the modeling. WH contributed to the modeling and the creation of the web application. TH contributed to the design of the study. SYW and HH performed manual reviews of the notes. SYW wrote the first draft of the manuscript. All authors contributed to the interpretation and subsequent edits of the manuscript. SYW is the guarantor.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

None.

Funding sources

Research to Prevent Blindness Career Development Award (SYW); National Institutes of Health National Eye Institute 1K23EY03263501 (SYW); Research to Prevent Blindness unrestricted departmental funds (SYW; WH); National Eye Institutes P30-EY026877 (SYW; WH).

The data underlying this article cannot be shared publicly due to the use of clinical notes in the study, which include many protected health identifiers containing patient information.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104864.

References

[1] J.A. Fries, E. Steinberg, S. Khattar, S.L. Fleming, J. Posada, A. Callahan, N.H. Shah, Ontology-driven weak supervision for clinical entity classification in electronic health records, Nat. Commun. 12 (1) (2021), https://doi.org/10.1038/s41467-021-22328-4.

[2] R.F. Sarmiento, F. Dernoncourt, Improving Patient Cohort Identification Using Natural Language Processing, in: Secondary Analysis of Electronic Health Records. Springer, Cham (CH), 2016.

[3] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention Is All You Need. arXiv [cs.CL]. 2017. http://arxiv.org/abs/1706.03762.

[4] K. Huang, J. Altosaar, R. Ranganath, ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv [cs.CL]. 2019. http://arxiv.org/abs/1904.05342.

[5] Alsentzer E, Murphy JR, Boag W, *et al.* Publicly Available Clinical BERT Embeddings. arXiv [cs.CL]. 2019.http://arxiv.org/abs/1904.03323.

[6] Y. Peng, S. Yan, Z. Lu, Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv [cs.CL]. 2019. http://arxiv.org/abs/1906.05474.

[7] J. Lee, W. Yoon, S. Kim, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[8] V. Sanh, L. Debut, J. Chaumond, et al., DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv [cs.CL]. 2019. http://arxiv.org/abs/1910.01108.

[9] M. Abadeer, 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. Online: Association for Computational Linguistics, 2020, 158–167.

[10] D.M. Baughman, G.L. Su, I. Tsui, C.S. Lee, A.Y. Lee, 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. 6 (2) (2017) 2, https://doi.org/10.1167/tvst.6.2.2.

[11] H.J. Lowe, T.A. Ferris, P.M. Hernandez, et al., STRIDE–An integrated standards-based translational research informatics platform, AMIA Annu. Symp. Proc. 2009 (2009) 391–395.

[12] *nltk*. Github https://github.com/nltk/nltk (accessed 2 Jun 2021).

[13] Wikipedia contributors. Inside–outside–beginning (tagging). Wikipedia, The Free Encyclopedia. 2020. https://en.wikipedia.org/w/index.php?title=Inside%E2%80%93outside%E2%80%93beginning_(tagging)&oldid=958799045 (accessed 1 Mar 2021).

[14] L.A. Ramshaw, M.P. Marcus, Text Chunking using Transformation-Based Learning. arXiv [cmp-lg], 1995. http://arxiv.org/abs/cmp-lg/9505040.

[15] T. Wolf, L. Debut, V. Sanh, et al., HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv [cs.CL]. 2019. http://arxiv.org/abs/1910.03771.

[16] seqeval. Github https://github.com/chakki-works/seqeval (accessed 1 Mar 2021).

[17] D. Nouvel, M. Ehrmann, S. Rosset, Evaluating named entity recognition. Named Entities for Computational Linguistics, (2016) 111–129. doi:10.1002/9781119268567.ch6.

[18] N. Perera, M. Dehmer, F. Emmert-Streib, Named Entity Recognition and Relation Detection for Biomedical Information Extraction, Front. Cell Dev. Biol. 8 (2020) 673.

[19] S.Y. Wang, J. Huang, S. Tao, eyelovedata/oph-notes-ner-slefe: v1.0.0. 2022. doi: 10.5281/zenodo.6977464.

[20] D. Mitry, T. Peto, S. Hayat, et al., Crowdsourcing as a screening tool to detect clinical features of glaucomatous optic neuropathy from digital photography, PLoS One 10 (2015) e0117401.

[21] X. Wang, L.I. Mudie, M. Baskaran, C.-Y. Cheng, W.L. Alward, D.S. Friedman, C. J. Brady, Crowdsourcing to Evaluate Fundus Photographs for the Presence of Glaucoma, J. Glaucoma 26 (6) (2017) 505–510.

[22] X. Wang, L. Mudie, C.J. Brady, Crowdsourcing: an overview and applications to ophthalmology, Curr. Opin. Ophthalmol. 27 (3) (2016) 256–261.

[23] T. Hernandez-Boussard, K.L. Monda, B.C. Crespo, D. Riskin, Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies, J. Am. Med. Inform. Assoc. 26 (11) (2019) 1189–1194.