Identify Diabetic Retinopathy-related Clinical Concepts and Their Attributes Using Transformer-based Natural Language Processing Methods

Zehao Yu[1], Xi Yang[1], Gianna L Sweeting[2], Yinghan Ma[1], Skylar E. Stolte[2], Ruogu Fang[2], Yonghui Wu[1§]

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, US

[2]Department of Biomedical Engineering, College of Engineering, University of Florida, Gainesville, Florida, US

[§]Corresponding author:

Yonghui Wu

Department of Health Outcomes and Biomedical Informatics

College of Medicine

University of Florida

Clinical and Translational Research Building, 2004 Mowry Road, PO Box 100177, Gainesville, FL, USA, 32610

Email address:

Zehao Yu – zehao.yu@ufl.edu; Xi Yang - alexgre@ufl.edu; Gianna L Sweeting - sweetinggianna@ufl.edu; Yinghan Ma - ma.yinghan@ufl.edu; Skylar E. Stolte - skylastolte4444@ufl.edu; Ruogu Fang - ruogu.fang@bme.ufl.edu; Yonghui Wu - yonghui.wu@ufl.edu

Abstract

Background

Diabetic retinopathy (DR) is a leading cause of blindness in American adults. If detected early, DR can be treated to preventing further damage causing blindness, therefore, early detection is very important for the treatment of DR. There is an increasing interest in developing AI technologies to help early detection of DR using electronic health records (EHR). The detailed diagnoses information documented in fundus image reports is a valuable resource that could help detect lesions from the medical image, thus helping early detection of DR. However, most studies for AI-based DR diagnoses are purely based on medical images; there is limited study to explore information captured in the free text image reports.

Methods

In this study, we examined two state-of-the-art transformer-based natural language processing (NLP) models, including BERT and RoBERTa, compared them with a recurrent neural network implemented using Long short-term memory (LSTM) to extract DR-related concepts from clinical narratives. We identified four different categories of DR-related clinical concepts including lesions, eye parts, laterality, and severity, developed annotation guidelines, annotated a DR-corpus of 536 image reports, and developed transformer-based NLP models for clinical concept extraction and relation extraction. We also examined the relation extraction under two settings including 'gold-standard' setting - where gold-standard concepts were used – and end-to-end setting.

Results

For concept extraction, the BERT model pretrained with the MIMIC III dataset achieve the best performance (0.9503 and 0.9645 for strict/lenient evaluation). For relation extraction, BERT

model pretrained using general English text achieved the best strict/lenient F1-score of 0.9316. The end-to-end system, BERT_general_e2e, achieved the best strict/lenient F1-score of 0.8578 and 0.8881, respectively. Another end-to-end system based on the RoBERTa architecture, RoBERTa_general_e2e, also achieved the same performance as BERT_general_e2e in strict scores.

Conclusions

This study demonstrated the efficiency of transformer-based NLP models for clinical concept extraction and relation extraction. Our results show that it's necessary to pretrain transformer models using clinical text to optimize the performance for clinical concept extraction. Whereas, for relation extraction, transformers pretrained using general English text perform better.

Keywords: Diabetic retinopathy, Natural language processing, named entity recognition, deep learning, relation extraction

Background

Diabetic Retinopathy (DR), a common complication of diabetes, is the leading cause of blindness in American adults and the fastest growing disease threatening nearly 415 million diabetic patients worldwide [1] [2]. This disease may cause no symptoms or only mild vision problems but eventually, can cause blindness. With professional eye imaging devices such as fundus cameras or Optical Coherence Tomography (OCT) scanners, most vision-threatening diseases can be curable if detected early [3]. Therefore, early detection is very important for effective treatment of DR. Recent development of Artificial Intelligence (AI) technology greatly improved the autonomous DR diagnosis systems including the referral system from Google AI and the FDA-approved iDx-DR, which make the early detection of vision-threatening diseases from a low-cost mobile camera available.

Electronic Health Records (EHR) have been increasingly implemented at US hospitals. Huge amounts of longitudinal patient data have been accumulated and are available electronically in structured tables, narrative text, and images. There is an increasing need for multimodal learning methods to link different data sources for clinical and translational studies. Recent emerging AI technologies, especially deep learning (DL) algorithms, have greatly improved the performance of automated vision-disease diagnoses systems based on EHR data. These AI systems for vision-disease diagnoses are usually developed using supervised machine learning models with medical images. The supervised machine learning models require annotated images, where the annotator have to manually label the region with lesions from images. In fact, the physicians have reviewed these medical images and documented detailed diagnosis, symptoms, and other critical observations in image reports, which could be a valuable resource to help annotators label images or serve as independent text features for lesion detection from medical images. There are

increasing numbers of clinical studies utilizing clinical narratives [4][5][6][7]. As the emergence of precision medicine, more and more studies look into clinical narratives to generate a more complete picture of patients to better assess health outcomes [8].

Natural language processing (NLP) is the key technology to extract patient information from clinical narratives to support various downstream clinical studies. Many NLP methods and systems have been developed to extract various types of information from clinical narratives. The clinical NLP community has organized a number of open challenges to advance information extraction from clinical narratives. Most state-of-the-art NLP methods for information extraction are based on supervised machine learning methods. The supervised machine learning models approach the information extraction as a two-stage pipeline, which typically include a clinical concept extraction (or named entity recognition [NER]) module to identify critical concepts (e.g., diseases, medications) and a relation extraction module to link attributes (e.g., negations, disease severity) to the concepts. For concept extraction, a number of NLP models have been developed to first identify clinical concepts and their attributes and then classify thenm into predefined semantic categories (e.g., diseases, medications). Relation extraction aims to establish semantic connections between extracted concepts and their attributes. Recently, transformer-based NLP models, built solely with a self-attention mechanism, outperformed other models and became state-of-the-art solution for information extraction from clinical narratives. For example, Peng et al. [9] proposed a BERT-based model for relation extraction; Dat et al. [7] proposed an end-to-end NLP model for relation and entity recognition in general English. However, the clinical text data is rarely used for developing AI systems for diagnosing DR and most studies on DR focused on medical images and structured EHRs. For example, Wong et al.[10] proposed a three-layer feed-forward neural network to detect the microaneurysms and hemorrhage from medical images; Imani

et al. [11] applied morphological component analysis to detect the exudation and blood vessel; Sun et al. [12] proposed a machine learning model to diagnose potential DR in patients using structured EHR data. There are studies exploring clinical narratives for text classification and computable phenotyping of DR. For example, Yang et al. [4] examined deep learning models to identify progress notes related to diabetes; Jin et al. [13] developed an NLP System to detect hypoglycemia-related events; Wu et al. [14] proposed a rule-based NLP system to help identify DR patients using clinical narratives. To the best of our knowledge, there are limited studies applying state-of-the-art transformer-based NLP models to extract DR-related clinical concepts from clinical narratives.

In this study, we identified patients diagnosed with DR at UF Health and collected their image reports, developed annotation guidelines and annotated a corpus for DR-related concept extraction, developed transformer-based NLP methods to extract DR-related clinical concepts that could help lesion detection from medical images. We systematically examined two state-of-the-art transformer-based NLP models for DR-related concept extraction and relation extraction from fundus image reports. We also developed end-to-end systems to detect DR-related concepts as well as their attributes in a unified system.

Methods

Data sets

We identified 155 patients diagnosed with diabetic retinopathy and collected a total number of 536 fundus image reports from them at the University of Florida (UF) Health. Then, we developed initial annotation guidelines through a collaboration of clinicians specialized in DR treatment, computer image experts (RF, SES, GLS), and NLP experts (YW, XY, ZY). Then, we recruited two annotators (YM, GLS) and conducted training sessions to help annotators get familiar with guidelines. We further improved the initial guidelines using several training sessions. After the

annotators achieved a good inter-annotator agreement score calculated using Cohen's Kappa [15] we conducted 3 rounds of annotation and finished the annotation of 536 notes. The first round (40 reports) was double-annotated to assess inter-annotator agreement. After each round of annotation, we discussed the discrepancies in group meetings, updated the annotation guidelines, and revised the annotations as needed.

DR-related concepts

There are many DR-related clinical concepts documented in the image reports such as diagnoses, treatments, and medications. As our goal is to extract DR concepts that can potentially help lesion detection from medical images, we identified four different categories of concepts, including lesions, eye parts, laterality, and severity. By definition, a lesion is a region in an organ or tissue which has suffered damage through injury or disease. In this study, we are particularly interested in lesions only associated with diabetic retinopathy (lesion occurred within the eye). Lesions that occurred in other organs were not be annotated. We also referred to the existing vocabulary of lesions [14], [15], and domain experts' knowledge to develop the annotation guidelines. When annotating a lesion, we asked annotators to annotate the associated eye-part, laterality, severity entity as well.

Annotation tool

We used the brat rapid annotation tool [18] for annotation. Fig. 1. shows an example of a DR-related lesion concept and the identified eye part, laterality, and severity.

Figure 1.

NLP methods

We adopted a standard two-stage NLP pipeline, including a clinical concept extraction module to detect DR-related concepts and their attributes and a relation extraction module to link the attributes to the corresponding concepts. For concept extraction, we used Long short-term memory (LSTM) model as a baseline and explored two state-of-the-art transformer-based NLP methods, including Bidirectional Encoder Representations from Transformers (BERT) [19] and Robustly optimized BERT approach (RoBERTa)[20] as they showed better performance in our previous study [21]. BERT is a bidirectional transformer-based NLP model based on masked language modeling (MLM) and uses next-sentence prediction (NSP) to learn representations from text. RoBERTa is a transformer-based language model shared the same architecture as BERT but pretrained with a dynamic MLM where masking patterns were generated during the training with different random seeds. We explored the LSTM model using Tensorflow. For transformer-based NLP models, we used the implementations from our clinical transformer package [21] based on the transformer architectures from the HuggingFace [22] in PyTorch [23]. For relation extraction task, we used the implementations from our clinical relation extraction with transformer package [24] based on the transformer architectures. Similar with the concept extraction task, we explored two state-of-the-art transformer-based NLP methods, including BERT and RoBERTa. As shown in Fig. 1, most relations between concepts occurred in the same sentence. Thus, we implemented heuristic rules to only consider two concepts occurring in the same as a candidate pair for relation classification.

For the LSTM model, following previous study on clinical concept extraction[25], we explored general models (LSTM_general) trained using English corpus using fastText [26] and compared the general models with clinical models (LSTM_clinic) trained using clinical notes from the Medical Information Mart for Intensive Care III (MIMIC-III) with the fastText algorithm. For

Transformer models, we used the 'base' setting in this study. Following our previous studies [21, 27, 28] on clinical transformers, we also examined pre-trained transformers from general English corpus (denoted as '_general', e.g., 'BERT_general') and clinical transformers pre-trained using clinical notes from the MIMIC-III database [29] (denoted as '_mimic', e.g., 'BERT_mimic'). We applied the default tokenizer in each model (e.g. wordpiece[30] in BERT and Byte-Pair Encoding[31] in RoBERTa) and adopted the default parameters optimized in our clinical transformer package and clinical relation extraction with transformer package [21] [24]. For relation extraction, we examined the transformer-based models under two settings, including (1) a pure relation extraction task where we assume that all concepts and their attributes are known and we only focus on how to identify the candidate pairs and classifier them into predefined categories, and (2) an end-to-end task to first identify the concepts and their attributes and then identify the relations (denoted as 'e2e'). For the end-to-end system, we applied the best model in concept extraction (BERT_mimic model) to generate candidate pairs and examined transformer models for relation classification.

Evaluation

We evaluated annotation agreement using Cohen's Kappa, κ, coefficient, where higher κ denotes annotator agreement. We used both strict (i.e., the beginning and end boundaries of a concept have to be exactly the same with gold-standard annotation) and lenient precision, recall, and F1-score to evaluate our NLP systems for concept extraction. Precision is defined as (the number of predicted concepts correctly identified by the NLP system) / (total number of concepts identified by NLP); recall is defined as (the number of predicted concepts correctly identified by the NLP system) / (total number of concepts annotated by experts); F1-score is defined as "(2*precision*recall)/(precision+recall)". We used the micro average to calculate the overall score.

Results

Table 1.

Two annotators annotated a total number of 4,782 DR-related concepts from 536 reports. The inter-annotator agreement measured by token level kappa score with 40 overlapped clinical notes was 0.74, indicating the two annotators have a reasonable agreement. We randomly divided the dataset into a training set and a test set with an 8:2 ratio. Table 1 shows the distribution of notes and DR-related concepts in the training and test set. We used the training set to develop transformer-based NLP models and used the test set for evaluation.

Table 2.

Table 2 compares six different NLP methods in extracting DR-related concepts from fundus image reports. All six methods performed well for concept extraction. The two transformer-based models outperformed the baseline LSTM model. Among four transformer-based models, the models pretrained using clinical notes from the MIMIC-III database outperformed their corresponding models pretrained using general English corpora. Among the two transformer-based NLP models trained using clinical text, the BERT_mimic model achieved the best strict/lenient F1-score of 0.9503 and 0.9645 on the test set, respectively. Table 3 shows the detailed performance for each of the four DR-related categories for the best NER model based on BERT. The BERT_mimic achieved lenient F1-scores over 0.95 for lesion, severity, and laterality, where the performance for detecting lesion is the best, which has a strict/lenient F1-score of 0.9565 and 0.9750, respectively; the performance for eye part category is relatively low with F1-score of 0.75.

Table 3.

Table 4 compares the two transformer-based NLP models for relation extraction under a gold-standard concept setting and an end-to-end setting. In the end-to-end systems, we applied the the best model for concept extraction – the BERT_mimic model. Using gold-standard concepts, the BERT_general achieved the best lenient/strict F1-scores of 0.9316. For the end-to-end setting, both BERT_general model and RoBERTa_general model achieved the best performance of 0.8578 using the strict evaluation. The BERT_general model achieved the best lenient F1-scores of 0.8881 under the end-to-end setting.

Table 4.

Conclusion and discussion

Identify DR-related concepts is a critical step to leverage clinical narratives for lesion detection from the medical image. In this study, we developed annotation guidelines to annotate DR-related concepts from fundus image reports, annotated a corpus of 536 image reports with four categories of clinical concepts, and examined two state-of-the-art transformer-based NLP models for detecting DR-related concepts and relations. For concept extraction, three out of four transformer-based models achieved better performance than the baseline model, except for the BERT_general model. The BERT model pretrained with the MIMIC III dataset achieved the best lenient F1-score of 0.9645. From Table 3, we noticed that the best model BERT_mimic achieved a good performance for lesion, severity, and laterality concepts, whereas, the performance for the eye part concept is relatively lower. One potential reason for the low performance for eye part concepts is there is limited number of concepts annotated compared with other categories. The transformer models pretrained using clinical text from the MIMIC III outperformed transformer models pretrained using general English corpora, which is consistent with findings reported in work [32,

33]. Similar to other clinical concept extraction tasks, fine-tuning the pre-trained transformers can further help improve the performance of extracting DR-related concepts.

We further link the severity, laterality, and eye part concepts to the corresponding lesion concept using relation extraction. The BERT_general model achieved the best strict/lenient scores of 0.8578 and 0.8881 for both settings, respectively. The RoBERTa_general also achieved the same performance as BERT_general in the strict evaluation score as a tie. Overall, the performance difference between the two transformer-based models in the end-to-end setting is not that significant with the setting using gold-standard concepts. It's not surprising to see that the performances for end-to-end systems are lower (~ 8% lower in strict evaluation and ~ 5% lower in lenient evaluation) than pure relation extraction using gold-standard concepts.

This study has limitations. The dataset we developed in this study is relatively clean without complex situations for relation extraction. For example, most of the relations are located at the same sentence. As the ultimate goal is to leverage the clinical narratives to help lesion detection from medical images, we plan to develop multimodal visual-text learning models to combine clinical text and medical images for early detection of DR in future studies.

List of abbreviations

DR	Diabetic retinopathy
NLP	Natural Language Processing
NER	Named Entity Recognition
EHR	Electronic Health Records
DL	Deep learning

LSTM	Long-Short Term Memory
UF	University of Florida
OCT	Optical Coherence Tomography
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly optimized BERT approach
NLM	Masked language modeling
NSP	Next-sentence prediction
MIMIC-III	Medical Information Mart for Intensive Care III

Declaration

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data is not publicly now.

Funding

This project was partially supported by a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2018C3-14754), a grant from National Institute on Aging 1R56AG 069880, and a SEED Grant from the UF Informatics Institute (00129436).

Acknowledgements

The authors would like to thank the NVIDIA Corporation with the donation of the GPUs used for this research.

Authors' contributions

ZY, XY, RF and YW were responsible for the overall design, development, and evaluation of this study. SES collected the data used in this study, ZY conducted the experiments and data analysis, XY was involved in the results analysis, too. YM and GLS annotated the concepts and relations. ZY and YW did the initial drafts and revisions of the manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

References

1. Bourne RRA, Stevens GA, White RA, Smith JL, Flaxman SR, Price H, et al. Causes of vision loss worldwide, 1990–2010: a systematic analysis. The Lancet Global Health. 2013;1:e339–49.

2. Mohamed Q, Gillies MC, Wong TY. Management of Diabetic Retinopathy: A Systematic Review. JAMA. 2007;298:902.

3. Gao Z, Li J, Guo J, Chen Y, Yi Z, Zhong J. Diagnosis of Diabetic Retinopathy Using Deep Neural Networks. IEEE Access. 2019;7:3360–70.

4. Yang B, Wright A. Development of deep learning algorithms to categorize free-text notes pertaining to diabetes: convolution neural networks achieve higher accuracy than support vector machines. arXiv:180905814 [cs, stat]. 2018. http://arxiv.org/abs/1809.05814. Accessed 21 Apr 2021.

5. Bucher BT, Shi J, Pettit RJ, Ferraro J, Chapman WW, Gundlapalli A. Determination of Marital Status of Patients from Structured and Unstructured Electronic Healthcare Data. AMIA Annu Symp Proc. 2020;2019:267–74.

6. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. J Am Med Inform Assoc. 2019;26:1163–71.

7. Nguyen DQ, Verspoor K. End-to-end neural relation extraction using deep biaffine attention. arXiv:181211275 [cs]. 2019;11437:729–38.

8. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. Journal of Biomedical Informatics. 2015;58:S128–32.

9. Shi P, Lin J. Simple BERT Models for Relation Extraction and Semantic Role Labeling. arXiv:190405255 [cs]. 2019. http://arxiv.org/abs/1904.05255. Accessed 22 Jul 2021.

10. Yun WL, Rajendra Acharya U, Venkatesh YV, Chee C, Min LC, Ng EYK. Identification of different stages of diabetic retinopathy using retinal optical images. Information Sciences. 2008;178:106–21.

11. Imani E, Pourreza H-R, Banaee T. Fully automated diabetic retinopathy screening using morphological component analysis. Computerized Medical Imaging and Graphics. 2015;43:78–88.

12. Sun Y, Zhang D. Diagnosis and Analysis of Diabetic Retinopathy Based on Electronic Health Records. IEEE Access. 2019;7:86115–20.

13. Jin Y, Li F, Yu H. HYPE: A High Performing NLP System for Automatically Detecting Hypoglycemia Events from Electronic Health Record Notes. arXiv:181111945 [cs]. 2018. http://arxiv.org/abs/1811.11945. Accessed 22 Apr 2021.

14. Wu H, Wei Y, Shang Y, Shi W, Wang L, Li J, et al. iT2DMS: a Standard-Based Diabetic Disease Data Repository and its Pilot Experiment on Diabetic Retinopathy Phenotyping and Examination Results Integration. J Med Syst. 2018;42:131.

15. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22:276–82.

16. Duh EJ, Sun JK, Stitt AW. Diabetic retinopathy: current understanding, mechanisms, and treatment strategies. JCI Insight. 2. doi:10.1172/jci.insight.93751.

17. Wang W, Lo ACY. Diabetic Retinopathy: Pathophysiology and Treatments. International Journal of Molecular Sciences. 2018;19:1816.

18. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France:

Association for Computational Linguistics; 2012. p. 102–7. https://www.aclweb.org/anthology/E12-2021. Accessed 23 Apr 2021.

19. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs]. 2019. http://arxiv.org/abs/1810.04805. Accessed 24 Sep 2020.

20. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:190711692 [cs]. 2019. http://arxiv.org/abs/1907.11692. Accessed 5 Mar 2021.

21. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. Journal of the American Medical Informatics Association. 2020;27:1935–42.

22. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:191003771 [cs]. 2020. http://arxiv.org/abs/1910.03771. Accessed 5 Mar 2021.

23. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:191201703 [cs, stat]. 2019. http://arxiv.org/abs/1912.01703. Accessed 5 Mar 2021.

24. Yang X, Yu Z, Guo Y, Bian J, Wu Y. Clinical Relation Extraction Using Transformer-based Models. arXiv:210708957 [cs]. 2021. http://arxiv.org/abs/2107.08957. Accessed 21 Jul 2021.

25. Yang X, Lyu T, Li Q, Lee C-Y, Bian J, Hogan WR, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Medical Informatics and Decision Making. 2019;19:232.

26. Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T. FastText.zip: Compressing text classification models. arXiv:161203651 [cs]. 2016. http://arxiv.org/abs/1612.03651. Accessed 26 Jul 2021.

27. Yang X, Zhang H, He X, Bian J, Wu Y. Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models. JMIR Med Inform. 2020;8:e22982.

28. Yang X, He X, Zhang H, Ma Y, Bian J, Wu Y. Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models. JMIR Med Inform. 2020;8:e19735.

29. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.

30. Schuster M, Nakajima K. Japanese and Korean voice search. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2012. p. 5149–52.

31. Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. arXiv:150807909 [cs]. 2016. http://arxiv.org/abs/1508.07909. Accessed 28 Jul 2021.

32. Ji Z, Wei Q, Xu H. BERT-based Ranking for Biomedical Entity Normalization. AMIA Jt Summits Transl Sci Proc. 2020;2020:269–77.

33. He Y, Zhu Z, Zhang Y, Chen Q, Caverlee J. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. arXiv:201003746 [cs]. 2020. http://arxiv.org/abs/2010.03746. Accessed 22 Apr 2021.

Figure list

1. An example of brat annotation for DR.

Table 1. Concepts distributions for training and test.

	training set	**test set**	**total**
Total notes	391	145	536
Lesion	2,383	896	3,279
Laterality	1,280	485	1,765
Severity	579	249	828
Eye part	45	17	62
Total concepts	4,287	1,647	5,934

Table 2. Performance comparison for concept extraction

	Strict			Lenient		
	precision	**recall**	**F1 score**	**precision**	**recall**	**F1 score**
LSTM_general	0.9492	0.9186	0.9337	0.9630	0.9320	0.9472
LSTM_mimic	0.9464	0.8682	0.9056	0.9609	0.8810	0.9192
BERT_general	0.8885	0.9575	0.9217	0.9067	0.9739	0.9391
BERT_mimic	0.9486	0.952	**0.9503**	0.9642	0.9648	**0.9645**
RoBERTa_general	0.9248	0.9636	0.9438	0.9353	0.9739	0.9542
RoBERTa_mimic	0.9391	0.9551	0.947	0.9498	0.9654	0.9575

* Best F1 scores are highlighted in bold.

Table 3. Detailed performance for each concept category for BERT_mimic

	Strict			Lenient		
	precision	**recall**	**F1 score**	**precision**	**recall**	**F1 score**
Lesion	0.9555	0.9576	0.9565	0.9776	0.9743	0.976
Severity	0.9627	0.9317	0.9469	0.9668	0.9357	0.951
Eye part	0.8	0.7059	0.75	0.8	0.7059	0.75
Laterality	0.9339	0.9608	0.9472	0.9439	0.9711	0.9573
Overall	0.9486	0.952	0.9503	0.9642	0.9648	0.9645

Table 4. Performance comparison for relation extraction models

		strict			lenient		
Settings	**NLP Models**	**precision**	**recall**	**F1 score**	**precision**	**recall**	**F1 score**
Use gold-standard concepts	BERT_general	0.9199	0.9437	**0.9316**	0.9199	0.9437	**0.9316**
	RoBERTa_general	0.9024	**0.9574**	0.9291	0.9024	**0.9574**	0.9291
	BERT_MIMIC	**0.9254**	0.9254	0.9254	**0.9254**	0.9254	0.9254
	RoBERTa_MIMIC	0.9147	0.9467	0.9304	0.9147	0.9467	0.9304
End-to-end	BERT_general_e2e	**0.8397**	0.8767	**0.8578**	**0.8712**	0.9056	**0.8881**
	RoBERTa_general_e2e	0.8274	**0.8904**	**0.8578**	0.8565	**0.9178**	0.8861
	BERT_MIMIC_e2e	0.8282	0.8584	0.843	0.8584	0.8858	0.8719
	RoBERTa_MIMIC_e2e	0.8362	0.8782	0.8567	0.8688	0.9072	0.8876

* Best precision, recall, and F1 are highlighted in bold. The strict and lenient scores are identical for the 'gold-standard' settings as the gold-standard annotation for concepts and attributes were used.