Assessing Disease Severity in Cutaneous Lupus Patients Using Natural Language Processing

UT Southwestern Medical Center

Laura Wang, BA[1], Kuroush Nezafati, MD[2], Ruichen Rong, PhD[2], Andrew Park, MD[1], Jane Zhu, MD[1], Guanghua Xiao, PhD[2], Yang Xie, PhD[2], Donghan M. Yang, PhD[2], Benjamin F. Chong, MD[1]

[1]Department of Dermatology, University of Texas Southwestern Medical Center [2]Quantitative Biomedical Research Center, Peter O'Donnell School of Public Health, University of Texas Southwestern Medical Center

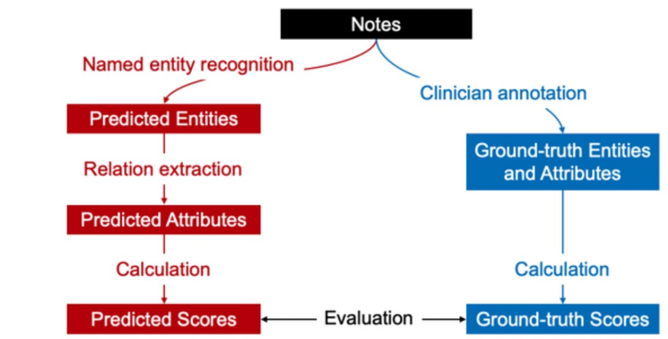
Background

- **Cutaneous lupus erythematous (CLE)** is an autoimmune skin disorder that manifests as inflammatory lesions in photosensitive areas.
- The **Cutaneous Lupus Erythematosus Disease Area and Severity Index (CLASI) score** measures disease activity and damage in CLE patients.
- There has been little work done previously using natural language processing (NLP) in dermatology to assess disease severity.
- **Objective:** To develop a NLP model that interprets physical examination (PE) documentation in CLE patients and computes disease severity scores in the form of CLASI activity and damage scores.

Methods

- Dataset was derived from 50 patients enrolled in the University of Texas Southwestern CLE registry.
  - **Training set:** 89 clinical exams of 24 patients were used to train the NLP model.
  - **Validation set:** 35 clinical exams of 26 patients were used to test the model's accuracy in prediction.
- An entity dictionary was provided rules for labeling vocabulary pertinent to CLASI scores within the PE note, as well as relationships between entities.
- The **BERT** (Bidirectional Encoder Representations from Transformers) **model**
  - Model was trained to predict all entities and relationships in the notes, based on which CLASI scores were calculated.
  - Model was then applied to the validation set. These scores were compared to the ground-truth CLASI scores based on human annotation (**Figure 1**).

Figure 1: Overall workflow



BERT was trained to extract entity and attributes and then calculate scores

Each note was annotated with ground truth entities and attributes. Ground-truth scores (CLASI scores) were provided by a dermatologist (BFC) who performed skin exams and scored patients in-person.

Results

Table 1: Cohort Characteristics

	Training (n = 24)	Validation (n = 26)	P-value
Age*	46.1 (39.5 – 57.5)	46.0 (39.7 – 56.1)	.68
Female	19 (79%)	24 (92%)	.18
Race			
Black	15 (63%)	16 (62%)	.94
Asian	0 (0%)	1 (4%)	.33
Caucasian	8 (33%)	6 (23%)	.42
Baseline CLASI Activity*	5.5 (1.74 – 12.5)	6 (2 – 15.8)	.68
Baseline CLASI Damage*	11 (4.5 – 15)	9.5 (3.5 – 11.8)	.43
Subtype			
DLE	20 (83%)	15 (58%)	<.05
SCLE	3 (13%)	5 (19%)	.52
TLE	1 (4%)	5 (19%)	.10
ACLE	0 (0%)	1 (4%)	.33

*Median of variables taken from the patients' earliest CLE-related visits within the registry

Figure 2: Performance in predicting individual scores

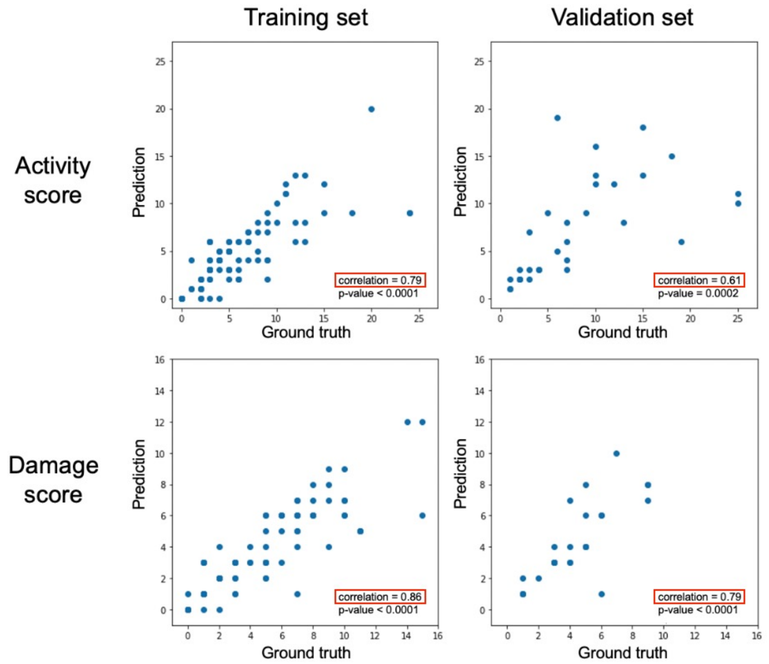


Figure 3: Performance in category prediction accuracy

Activity

Train accuracy: 0.90

Train	Pred mild	Pred severe
True mild	72	9
True severe	0	8

Validation accuracy: 0.90

Validation	Pred mild	Pred severe
True mild	19	2
True severe	1	9

Damage

Train accuracy: 0.87

Train	Pred mild	Pred severe
True mild	53	10
True severe	2	24

Validation accuracy: 0.85

Validation	Pred mild	Pred severe
True mild	17	1
True severe	3	6

Table 2: Performance in predicting score categories

	Accuracy*		Precision*		Recall*		F1*	
	Activity	Damage	Activity	Damage	Activity	Damage	Activity	Damage
Training	0.90	0.87	0.89	0.84	1.0	0.96	0.94	0.90
Validation	0.90	0.85	0.90	0.94	0.95	0.85	0.93	0.89

Discussion

- Our model predicted individual CLASI scores with high precision and recall, with better results on damage; not seen extensively in existing literature
- When predicting score severity categories, our model scored highly for accuracy, recall, precision, and F1 score for both training and validation sets

Limitations

- Single center study
- PE notes from a single clinician
- Notes may be subjective in terms of quality/style
- Newer language models such as GPT may perform stronger

Conclusions

- BERT-based NLP model can be trained to predict CLASI scores in CLE patients using PE notes.
- **Implication:** This study can significantly increase volume of real-world data available for CLE research by efficiently processing EHR
- **Future steps:** We will increase size and representation of the training set to improve accuracy and external validity of BERT's predictions.

Acknowledgements

We would like to thank the participants of the UTSW Cutaneous Lupus Erythematosus Registry for their contributions to lupus research. We would also like to thank our funding support including the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health (K23AR061441) (BC).