SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 1326.
https://doi.org/10.62617/mcb1326

Article

Weakly-supervised natural language processing with BERT-Clinical for automated lesion information extraction from free-text MRI reports in multiple sclerosis patients

Qiang Fang[1,*], Ryan J. Choo[2,3], Yuping Duan[4], Yuxia Duan[5], Hongming Chen[1,6], Yun Gao[1], Yunyan Zhang[2,3,7,*] Zhiqun Mao[8]

[1] School of Marine Engineering Equipment, Zhejiang Ocean University, Zhoushan 316022, China
[2] Departments of Radiology, University of Calgary, AB T2N 4N1, Canada
[3] Department of Clinical Neurosciences, University of Calgary, AB T2N 4N1, Canada
[4] Qingdao Central Hospital, University of Health and Rehabilitation Sciences, Qingdao 266042, China
[5] School of Physics and Electronics, Central South University, Changsha 410083, China
[6] College of Integrated Circuits, Zhejiang University, Hangzhou 310000, China
[7] Hotchkiss brain institute, University of Calgary, AB T2N 4N1, Canada
[8] Department of PET Imaging Center, Hunan Provincial People's Hospital, Changsha 410013, China
*** Corresponding authors:** Qiang Fang, happyqiangfang123@163.com; Yunyan Zhang, 895434290@qq.com

CITATION

Fang Q, Choo RJ, Duan Y, et al. Weakly-supervised natural language processing with BERT-Clinical for automated lesion information extraction from free-text MRI reports in multiple sclerosis patients. Molecular & Cellular Biomechanics. 2025; 22(4): 1326. https://doi.org/10.62617/mcb1326

ARTICLE INFO

Received: 8 January 2025
Accepted: 14 February 2025
Available online: 28 February 2025

Abstract: Purpose: To investigate how bidirectional encoder representations from transformers (BERT)-based models help extract treatment response information from free-text radiology reports. **Materials and methods:** This study involved 400 brain MRI reports from 115 participants with multiple sclerosis. New MRI lesion activity including new or enlarging T2 (newT2) and enhancing T1 (enhanceT1) lesions for assessing treatment responsiveness was identified using the named entity recognition technique along with BERT. Likewise, 2 other associated entities were also identified: the remaining brain MRI lesions (regT2), and lesion location. Report sentences containing any of the 4 entities were labeled for model development, totally 2568. Four recognized BERT models were investigated, each with conditional random field integrated for lesion versus location classification, trained using variable sample sizes (500–2000 sentences). Regularity was then applied for lesion subtyping. Model evaluation utilized a flexible F1 score, among others. **Results:** The Clinical-BERT performed the best. It achieved the best testing flexible F1 score of 0.721 in lesion and location classification, 0.741 in lesion only classification, and 0.771 in regT2 subtyping. With growing sample sizes, only Clinical-BERT performed increasingly better, which also had the best area under the curve of 0.741 in lesion classification at training using 2000 sentences. The PubMed-BERT achieved the best testing flexible F1 score of 0.857 in location only classification, and 0.846 and 0.657 in subtyping newT2 and enhanceT1, respectively. **Conclusion:** Based on a small sample size, our methods demonstrate the potential for extracting critical treatment-related information from free-text radiology reports, especially Clinical-BERT.

Keywords: named entity recognition; lesion; free-text reports; semi-supervised learning; conditional random field; BERT

Effective extraction of pertinent information from electronic health records is highly valuable for empowering healthcare [1]. Of particular importance is the ability to extract information from free-text documents involved in daily clinical practice such as radiology reports. In the context of neurological disorders such as multiple sclerosis (MS), magnetic resonance imaging (MRI) serves as a critical tool for both disease diagnosis and treatment [2]. For the latter, identifying the presence and nature of new

MRI lesion activity is essential as part of a well-recognized criterion known as No Evident Disease Activity (NEDA) for characterizing treatment response in MS [3]. Currently this identification task relies mainly on manual processes, which are time-consuming and prone to human errors. Therefore, the availability of a robust automatic method such as natural language processing (NLP) is highly desirable.

By enabling direct human-computer interactions, the NLP has shown considerable potential to handle various tasks in different fields including health [4]. Example tasks included information retrieval [5], text abstraction [6], and question-answering [7]. With emergence of large language models such as the bidirectional encoder representations from transformers (BERT), the capacity of NLP has grown substantially [8]. Specifically, equipped with current deep learning technologies and innovative self-attention mechanisms, BERT models have demonstrated promise in different studies such as disease subtyping [9], disease risk prediction [10], and information extraction [11]. Further, by leveraging intricate context and semantics representations, BERT has also enabled an improved accuracy in named entity recognition (NER), a crucial NLP task for identifying key elements in text [12]. Additionally, depending on the type of training samples used, there have been different promising variants of BERT, including BERT-base-cased/uncased [13], PubMed-BERT [14], and Clinical-BERT [15].

Nonetheless, how these models work with domain-specific applications such as real-world radiology reports are unclear, especially those with a small sample size.

The goal of this study was to implement customized BERT models capable of extracting treatment response information at the token level along with NER based on free-text MRI reports of MS participants. Through transfer learning, this study also aimed to investigate and compare different BERT variants, together with implementation of competitive NLP techniques including classification and regularity search. We hypothesized that the approaches were feasible for conducting the domain-specific tasks and that models trained with biomedical documents would perform better than those with general documents.

1. Materials and methods

1.1. Sample characteristics (radiology reports corpus)

This was a retrospective feasibility study comprised of 400 free-text brain MRI reports collected between 1 December 2016, and 31 March 2020, from a convenience sample of 115 persons with MS. All participants were under routine clinical care and were starting or switching to a new disease-modifying therapy as part of an ongoing clinical project. The mean (range) age was 42 (25–67) years, disability scores ranged from 0 to 5.5, and disease duration was 1–15 years; 65 participants were women. Each individual had 1 to 3 sequential MRI reports included, which were de-identified through a pseudo-anonymization process, followed by encryption of all study data. Importantly, the reports were collected from three different branches of Foothills Hospital, University of Calgary, located across the Alberta region. This multi-branch data collection strategy ensures diversity in the dataset, as the reports were generated by different radiologists, using varying imaging protocols, and serving diverse patient populations. This approach inherently addresses external validation concerns, as the

dataset reflects real-world variability and enhances the generalizability of our findings. This study was approved by the Institutional Ethics Board, and all participants provided written informed consent.

1.2. Terminology

There were four groups of lesion entities (**Table 1**). These were: new or enlarging lesions on T2-weighted MRI (newT2), enhancing lesions on post-contrast T1-weighted MRI (enhanceT1), other lesions on T2-weighted or T2-FLAIR MRI (regT2), and location of these lesions (location). The initial reports contained vocabulary and terms representing defined variables or a combination of variables.

Table 1. Terminology used to describe each entity category of lines and tubes.

Entity Category	Include Terms/key Words
Lesion	Lesions/no/without/
Location	supratentorial white matter
	posterior limb left internal capsule posterior fossa white matter/
Brain new/enlarging T2 lesion	new increased
	increasing
Brain enhancing T1 lesion	Enhancing enhancing larger
	enhanced
Brain regular T2 lesion	multiple lesions
	abnormal signal

Note: FLAIR: Fluid attenuated inversion recovery.

1.3. Data curation

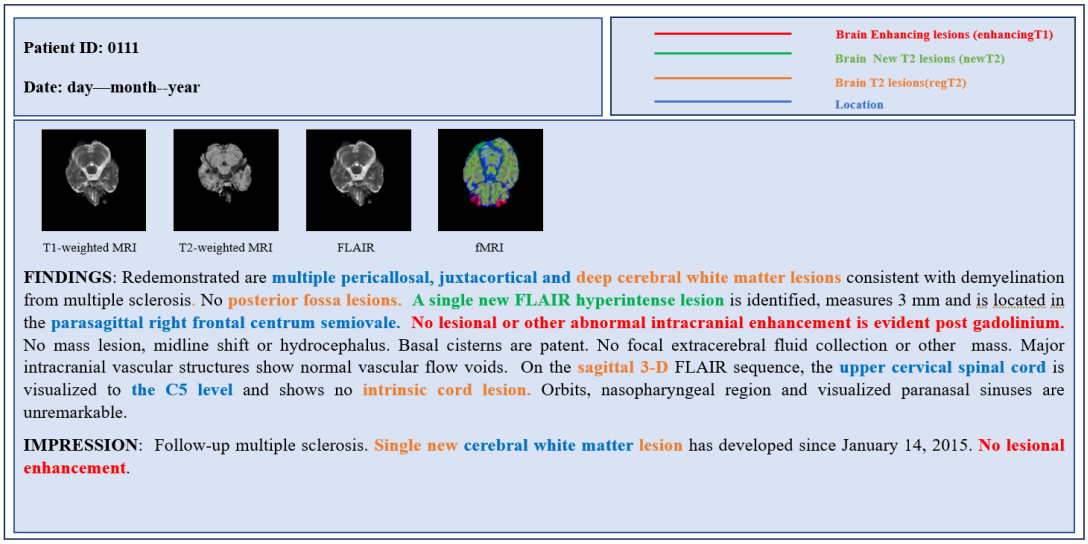


Figure 1. A data curation example. Shown is a free-text MRI report associated with this study following annotation using our color-coding system developed for the study. Top right shows the four defined entity categories and their affiliated colors, including three lesion subtypes: newT2, enhanceT1, and regT2, and Location of them. Bottom panel shows the main content of the report with color-highlights of the respective phrases and sentences containing each of the entity categories.

For all the free-text MRI reports, sentences containing any of the four entities defined above were annotated based on a color-coding system developed for this study, one color per entity type. This process was done firstly by a PhD researcher (QF) and then a senior radiology resident (RJC) for confirmation (**Figure 1**). The identified entity sentences were subsequently categorized by type (color) and divided randomly into 3 portions for model training (78%), validation (10%), and testing (12%).

1.4. Software and hardware

Model development and testing used algorithms implemented with PyTorch (version 1.7.1 + cu110) [16]. The workstation employed had Tesla T4 GPUs installed.

1.5. Model development

1.5.1. Architecture of domain-specific BERT models for NER

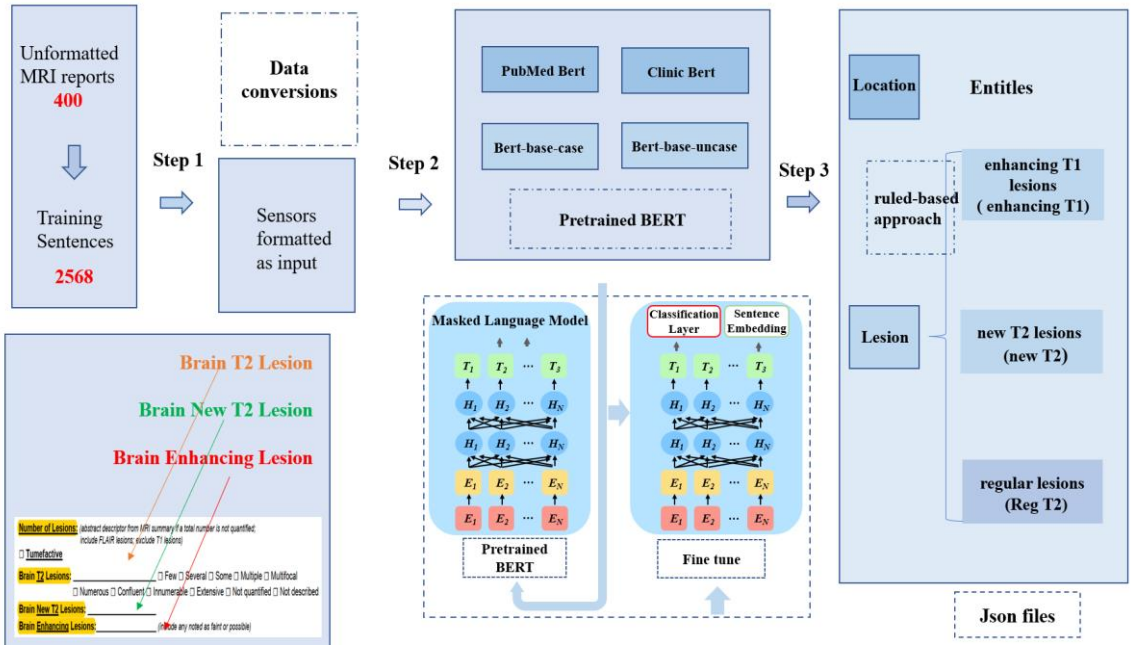


Figure 2. Overview of the study design. The initial steps (left panel) focus on data curation, including annotation of the free-text MRI reports into entity-containing sentences and conversion of the sentences into numeric tensors ready for modeling using our domain-specific architectures (middle panel). Each architecture is comprised of a pre-trained BERT network and a conditional random field (CRF) layer, which generate outputs of 2 main entity categories: lesion and location (right panel). Subsequently, the lesion entity is further divided into three subtypes: new or enlarging T2 lesions (newT2), enhancing T1 lesions (enhanceT1), and others (regT2), through regularity analysis enabled by our implementation of a rule-based mechanism.

Note: BERT = bidirectional encoder representations from transformers.

Based on pre-trained BERT models, this study developed four domain-specific architectures for NER: BERT-base-cased/uncased, PubMed-BERT, and Clinical-BERT. The former pair were smaller yet faster versions than the vanilla BERT and exhibited superior performance [17,18], with differences of the BERT-base-uncased being pre-trained on lowercased text by disregarding the original capitalization of words. The PubMed-BERT was pre-trained on English-language text from the

biomedical field such as PubMed abstracts, which employed a dynamic version of masked language modeling for effective fine-tuning on downstream tasks. The Clinical-BERT was pre-trained on clinical notes, which employed disentangled attention [19] for an automatic focus on distinct elements of the input data, as well as an improved mask decoder to enhance performance. To achieve NER, each BERT model was integrated with a classification component using the Conditional Random Field (CRF) approach (**Figure 2**).

1.5.2. Implementation

Using model-specific tokenizers, our categorized entity sentences were embedded into token level numerical data, which served as input to our domain-specific models. The pre-trained BERT generated high-dimensional tensors sized S × 768, which represented sequence length and embedding size, respectively. The CRF model produced token-wise tensors, each with 5 dimensions based on the BIESO tag scheme [20], with each dimension assigned a probability to represent a token. The final output of our model was the BIESO label probabilities that identified our main entities: lesion or location (**Figure 3**).

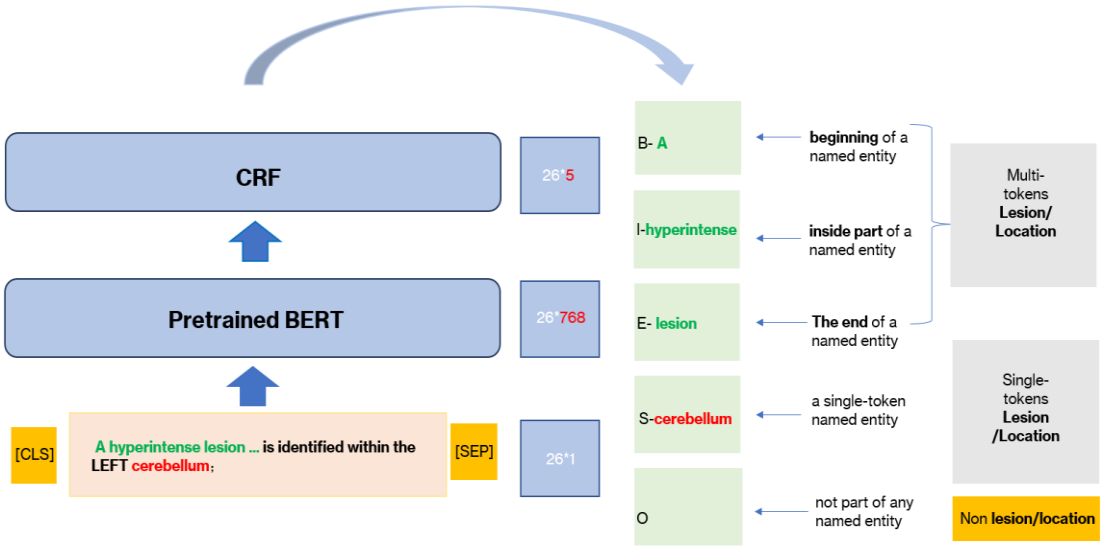


Figure 3. An example process of our domain-specific models. Shown as input is an example entity sentence (bottom left): A hyperintense lesion … is identified within the LEFT cerebellum along the posterior margin of the fourth ventricle which is not definitely appreciated on the prior study. Using a model-specific tokenizer, the sentence is converted into a 26 × 1 vector of token identifiers that is input to the corresponding BERT network. The latter generates a 26 × 768 information tensor, which is analysed by the CRF to generate label probabilities sized 26 × 5 following mapping to the B, I, E, S, and O tags (mid & right columns).

Note: CLS = Classification Token; SEP = Separator Token; BERT = bidirectional encoder representations from transformers.

Model optimization used the Adam optimizer [21] by minimizing an average loss [22], with a learning rate of 5×10^{-5} as recommended. Over training, each run consisted of 20 epochs. After each run, models having the lowest validation loss were selected for subsequent runs. This process was repeated 5 times for each of our domain-specific models with seed initializations set randomly per repeat to understand stability, where all parameters were unfrozen for fine-tuning for each model. In

addition, each model was additionally trained using different sample sizes, at 25% (500/2000), 50% (1000/2000), 75% (1500/2000), and 100% (2000/2000) of the available sentences, and validated accordingly to further investigate reliability.

1.6. Regularity analysis

We implemented a rule-based approach to further process the prediction results of our models on the lesion entity. The rules were defined based on the presence of vocabulary and phrases associated with each of the pre-defined lesion subtypes: newT2, enhanceT1, and regT2. When no keywords were detected for the first 2 categories after an initial round of lexical extraction, the entity was assigned to the third, regT2. This was done for each model, followed by confusion matrix construction per lesion subtype.

1.7. Model evaluation

For each domain-specific model, the model architecture with the lowest validation loss was carried over for testing on the held-out dataset. This study focused on 3 types of evaluation metrics recognized in this field: F1 score with customization (flexible F1), area under the receiver operating characteristic curve (AUC), and confusion matrix. The F1 score combined precision and recall [23]. Here our flexible F1 score aimed at assessing the most important information being identified. That is, a result was considered true positive if all key entities were correctly identified despite the missing of certain non-important tokens. Therefore, this metric could serve as a more practical evaluation of the NLP models than the default version. Further, our confusion matrices provided model assessment at different levels, ranging from main entities (lesion, location, and both combined) to lesion subtypes.

1.8. Statistical analysis

The mean and 95% confidence interval of AUC were calculated based on results from the testing set. The DeLong test [24] was used to compare AUC values between models. To account for multiple comparisons, the Benjamini-Hochberg method was applied. In addition, flexible F1 score was used to compare models trained using different sample sizes per architecture. All statistical analyses were conducted using Python, with $P \leq 0.05$ set as significance.

2. Results

2.1. The domain-specific models performed differently in identifying the main lesion and location entities

Based on all entity sentences available, our fine-tuned models achieved an AUC of approximately 0.562 to 0.712 in identifying the main lesion and location entities. The Clinic-BERT was the best in classifying the lesion as well as lesion and location combined entities, with an AUC of 0.712 and 0.700, respectively. The BERT-base-uncased model also achieved an AUC of 0.712 in classifying the combined entity, and the best AUC of 0.674 in classifying the location entity (**Figure 4**—top panels). In pair-wise comparison of AUC values, DeLong test showed that the Clinical-BERT

was better than PubMed-BERT in classifying both the lesion ($p = 0.044$) and combined ($p = 0.014$) entities. In contrast, both the BERT-base-cased ($p = 0.006$) and PubMed-BERT ($p = 0.025$) performed better than Clinic-BERT in distinguishing the location entity. The trend of difference between models was maintained after correction for multiple comparisons except the loss of significance, where Clinical-BERT showed the highest trend to be significant ($p = 0.055$) in classifying the combined entities as compared to PubMed-BERT (**Table 2**).

Table 2. Model comparisons by AUC through Delong test.

	Model 1	**AUC 1**	**Model 2**	**AUC 2**	***P* value**	**Adjust *P* value**
Lesion	BERT-base-cased	0.687	BERT-base-uncased	0687	0.124	0.241
	BERT-base-cased	0.687	Pub-med BERT	0.667	0.518	0.657
	BERT-base-cased	0.687	Clinical-BERT	0.700	0.004	0.098
	BERT-base-uncased	0.687	Pub-med BERT	0.667	0.225	0.090
	BERT-base-uncased	0.687	Clinical BERT	0.700	0.281	0.127
	Pub-med BERT	0.667	Clinical BERT	0.700	0.044	0.079
Location	BERT-base-cased	0.575	BERT-base-uncased	0.674	0.615	0.301
	BERT-base-cased	0.575	Pub-med BERT	0.619	0.358	0.333
	BERT-base-cased	0.575	Clinical-BERT	0.562	0.006	0.245
	BERT-base-uncased	0.674	Pub-med BERT	0.619	0.781	0.559
	BERT-base-uncased	0.674	Clinical BERT	0.562	0.918	0.705
	Pub-med BERT	0.619	Clinical BERT	0.562	0.025	0.132
Combined	BERT-base-cased	0.702	BERT-base-uncased	0.712	0.814	0.208
	BERT-base-cased	0.702	Pub-med BERT	0.673	0.075	0.112
	BERT-base-cased	0.702	Clinical-BERT	0.712	0.509	0.224
	BERT-base-uncased	0.712	Pub-med BERT	0.673	0.060	0.075
	BERT-base-uncased	0.712	Clinical BERT	0.712	0.332	0.326
	Pub-med BERT	0.673	Clinical BERT	0.712	0.014	0.055

Note: Bold font Indicates significant values where $p \leq 0.05$ or $p \leq 0.01$; italics indicates close to significance. AUC: area under the receiver operating characteristic curve.

Model assessment using our flexible F1 score showed similar results based on training using most (2000) entity sentences. The Clinical-BERT had the highest flexible F1 score of 0.74 in classifying the lesion entity, and among the high end with a score of 0.72 in classifying the combined entity. For most cases, the predicted results of this model closely resembled ground truth, with only minor discrepancies in identifying non-essential tokens (**Figure 4**). The PubMed- BERT and BERT-base-uncased performed the best in classifying the location and combined entities, respectively, with flexible F1 scores of ~0.86 and ~0.73.

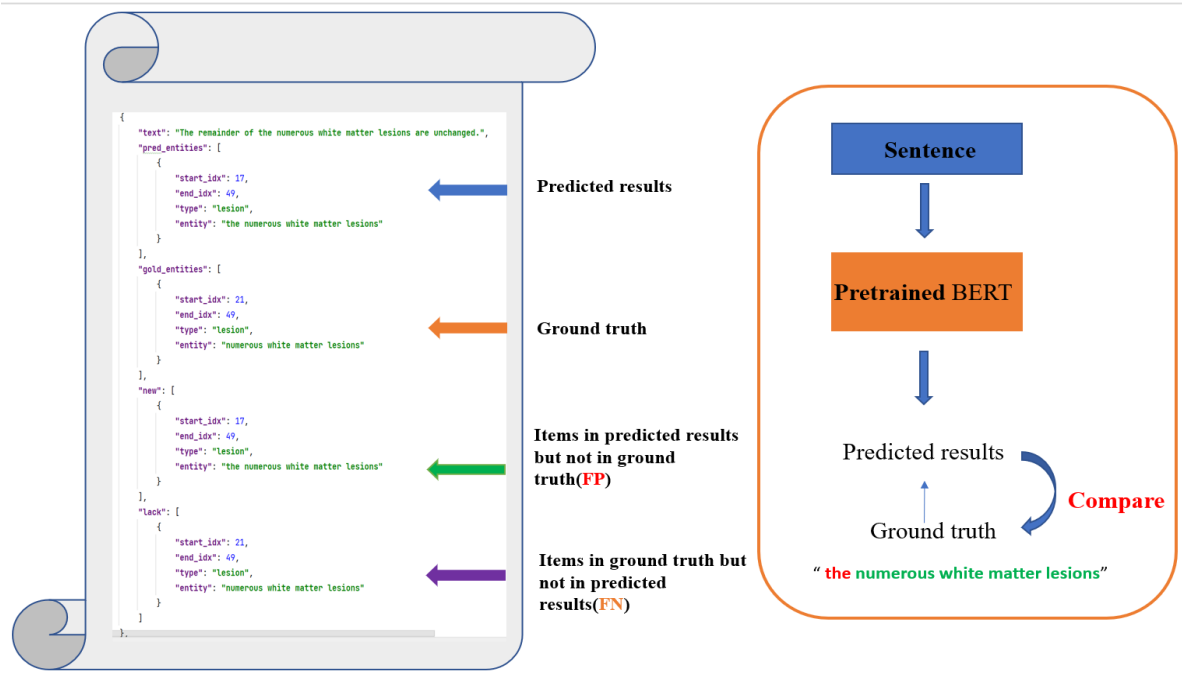


Figure 4. Example outputs and evaluation of our domain-specific models. Left panel shows a segment of results saved in a json file from our Clinical-BERT based on an example sentence: The reminder of the numerous white matter lesions are unchanged. The json segment includes four sections, with the first two representing predicted and gold (ground truth) entities, respectively. The third section known as "new" highlights items present in the predicted results but not in the ground truth (false positives, FP) such as the preposition 'the', while the fourth section referred to as "lack" indicates items present in the ground truth but not in the predicted results (false negative, FP); none presents here. Comparison between predicted and ground truth results (right panel) across the four sections allows us to generate the evaluation metrics, including our flexible F1 score.

Note: BERT = bidirectional encoder representations from transformers.

2.2. Regularity analysis permitted lesion subtyping

Based on the classification results of our main entities, regularity analysis allowed to further assess the performance of the models in characterizing lesion sub-categories. The BERT-base-uncased and PubMed-BERT models had a relatively higher performance than BERT-base-cased and Clinic-BERT models in subtyping the newT2 and enhanceT1 lesion entities. But in general, the performance of all models appeared to fluctuate or decrease as demonstrated by confusion matrices (**Figure 5**—bottom panels).

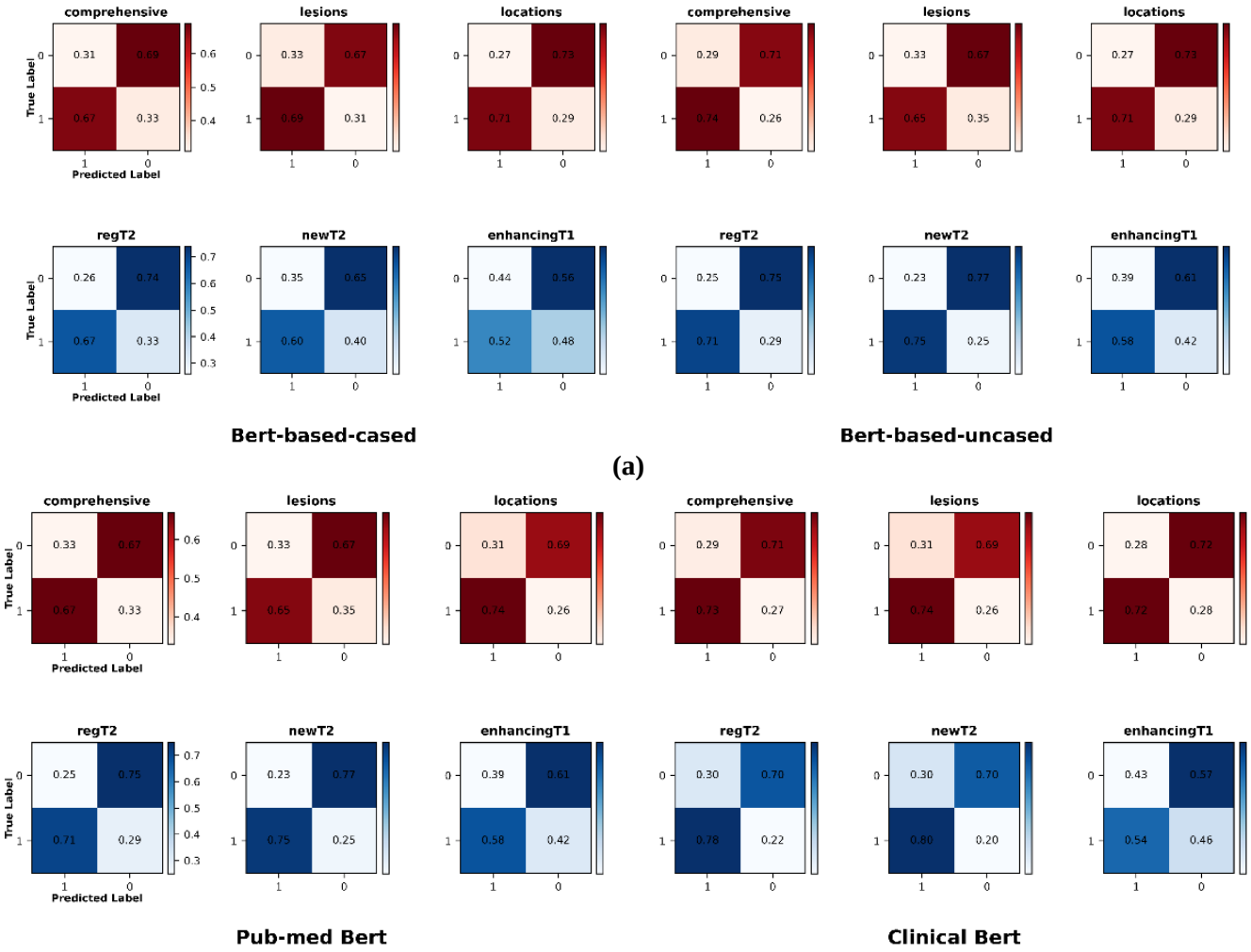


Figure 5. Confusion matrices for both main and lesion subtype entities based on each of our domain-specific models. For each model, shown on the top row are results for the main entities evaluated: Lesion, (lesion) Location, and both Combined; on the bottom row are results for the lesion subcategories: new or enlarging T2 lesions (newT2), enhancing T1 lesions (enhanceT1), and other lesions (regT2) following regularity analysis. For both ground truth (True label) and predicted results, '1' and '0' represent positive and negative labels, respectively. **(a)** Bert-based-cased(left), Bert-based-uncased(right); **(b)** Pub-med Bert(left), Clinical Bert(right).

Note: BERT = bidirectional encoder representations from transformers.

2.3. Domain-specific Clinical-BERT performed competitively and consistently better with training by a growing sample size

To assess the impact of sample size on model performance, we investigated four training schemes based on 500–2000 entity sentences. With only 500 (25%) sentences, all models showed an AUC of > 0.640. Specifically, the BERT-base-uncased model achieved the highest AUC of 0.725 in classifying the main lesion and location entities, followed by Clinical-BERT,

PubMed-BERT and Bert-base-uncased, which achieved the highest AUC of 0.700, 0.677, and 0.670, respectively. When training sample size increased to 1000 (50%), 1500 (75%), and 2000 (100%) sentences, only the AUC of Clinical-BERT gradually improved. Further, the Clinical-BERT achieved the highest AUC of 0.741

among all study models under the 2000 sentence training scheme (**Figure 6**).

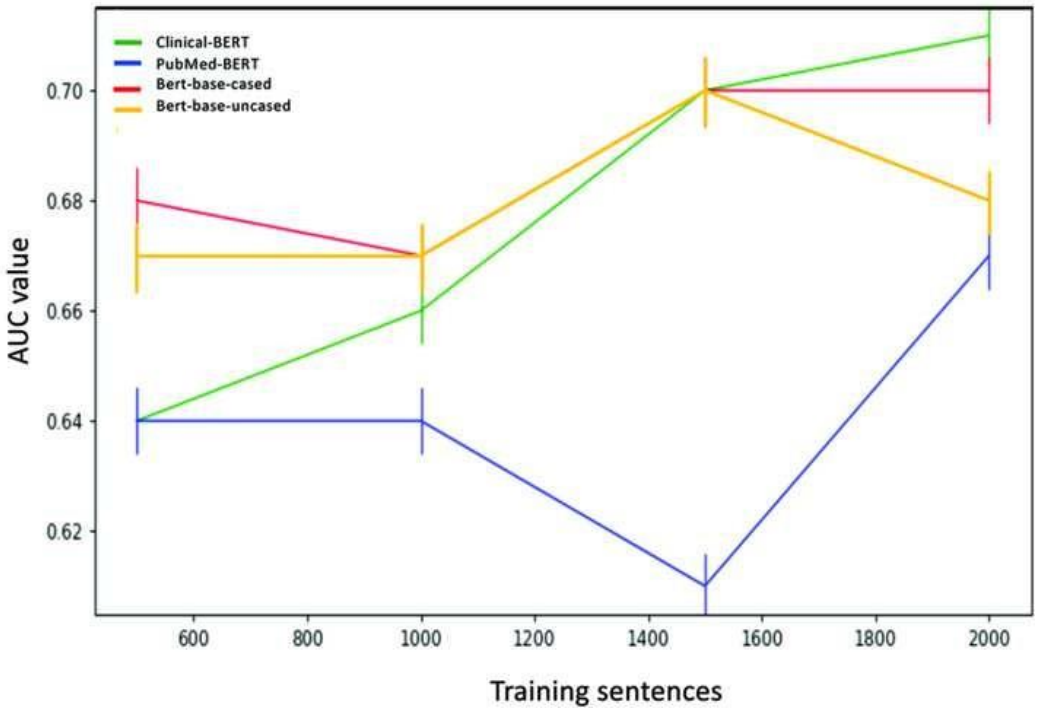


Figure 6. Dynamic AUC results of each domain-specific model investigated in the study with training using different sample sizes. Shown are mean and standard deviations. The sample sizes tested are 500 (25%), 1000 (50%), 1500 (75%), and 2000 (100%) entity sentences randomly chosen from the 2568 sentences available for this study. Each color line represents one type of domain-specific model.

Note: BERT = bidirectional encoder representations from transformers.

Table 3. Flexible F1 score of domain-specific models trained using different sample sizes of entity sentences.

Entity	Model	# (percent) Training sentences			
		500 (25%)	1000 (50%)	1500 (75%)	2000 (100%)
lesion	Bert-base-cased	0.653	0.722	0.730	0.726
	Bert-base-uncased	0.725	0.664	0.729	0.735
	PubMed-Bert	0.677	0.640	0.658	0.702
	Clinical Bert	0.700	0.728	0.731	0.741
location	Bert-base-cased	0.167	0.143	0.636	0.724
	Bert-base-uncased	0.133	0.105	0.680	0.546
	PubMed Bert	0.154	0.143	0.583	0.857
	Clinical Bert	0.133	0.154	0.740	0.714
Combined	Bert-base-cased	0.640	0.701	0.726	0.721
	Bert-base-uncased	0.706	0.649	0.717	0.726
	PubMed Bert	0.662	0.629	0.654	0.709
	Clinical Bert	0.682	0.701	0.717	0.721

Note: BERT = bidirectional encoder represents.

Based on the flexible F1 score, the Clinical-BERT again showed a persistently increasing performance with training using an increasingly larger sample size. For the

lesion entity category, the flexible F1 score of Clinical-BERT was 0.700 in training with only 500 entity sentences, and the score increased to 0.741 when training sentences increased from 2000. Similarly, for the combined entity, the Clinical-BERT also showed a high flexible F1 score of 0.682 under training with 500 sentences, and an increasingly higher score of up to 0.712 when training sample increased to 2000 sentences. No other model showed a consistent trend with gradual change of the sample sizes for any main entity (**Table 3**).

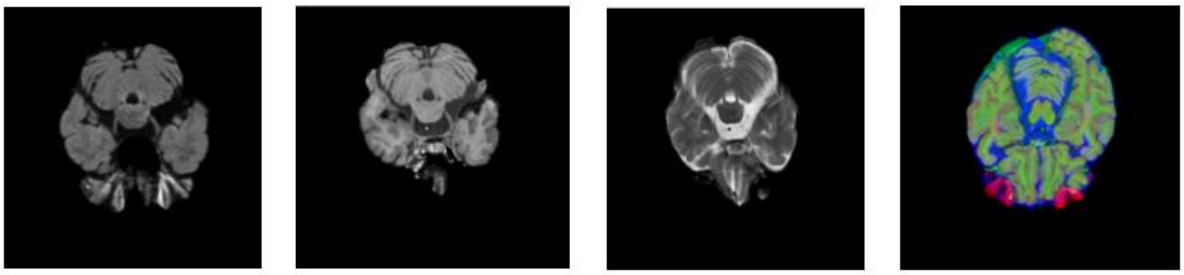
2.4. Time efficiency

With a similar number of parameters involved in the study models, the Clinical-BERT was the fastest. It took 1 (one) min and 26 s in training and validation on 2568 entity sentences inclusive, and 6.29 s in testing on 309 held-out sentences. The BERT-base-uncased model was the slowest, requiring 5 min and 45 s for training and validation, and 6.496 s for testing. The PubMed-BERT was the second fastest that required 2 min and 52 s over training and validation, and 5.912 s on testing (**Table 4**).

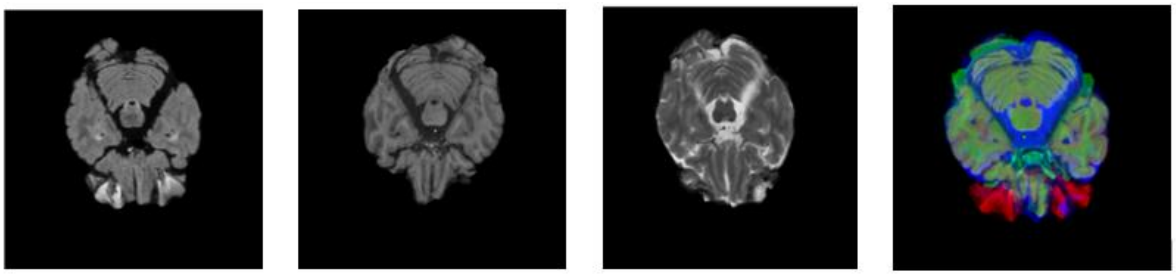
Table 4. Running time comparison between models based on 2000/2568 entity sentences.

Model	Training/Validation Time	Test time	Parameter Count
Bert-base-cased	5 min 44 s	5.768 s	111.94 M
Bert-base-uncased	5 min 45 s	6.496 s	110.77 M
PubMed BERT	2 min 52 s	5.912 s	111.94 M
Clinic-BERT	1 min 26 s	6.298 s	111.04 M

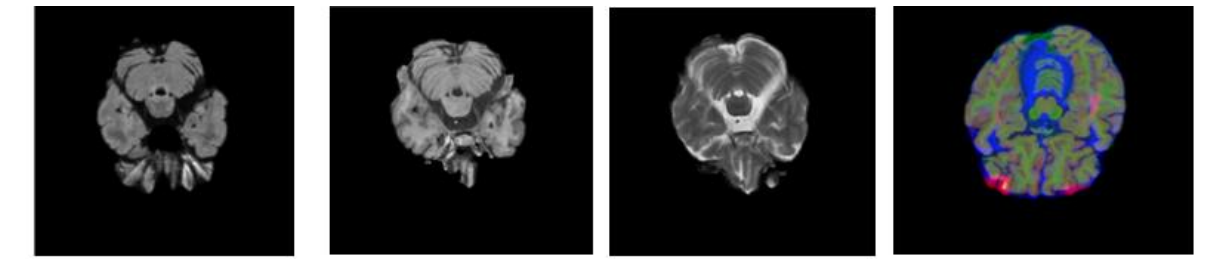
2.5. Representative cases for lesion extraction and data fusion in multiple sclerosis



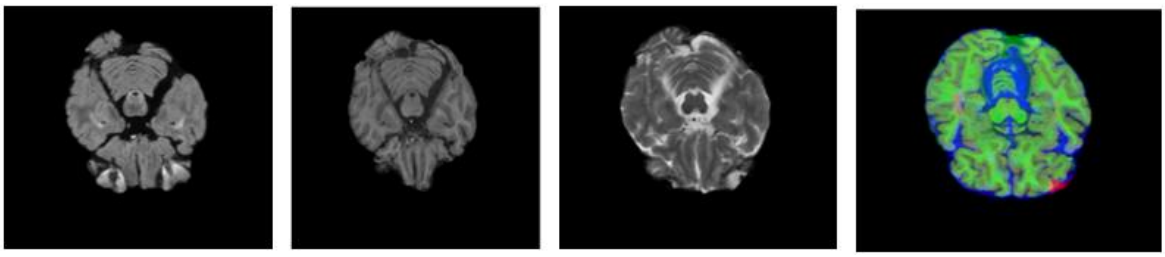
(a)



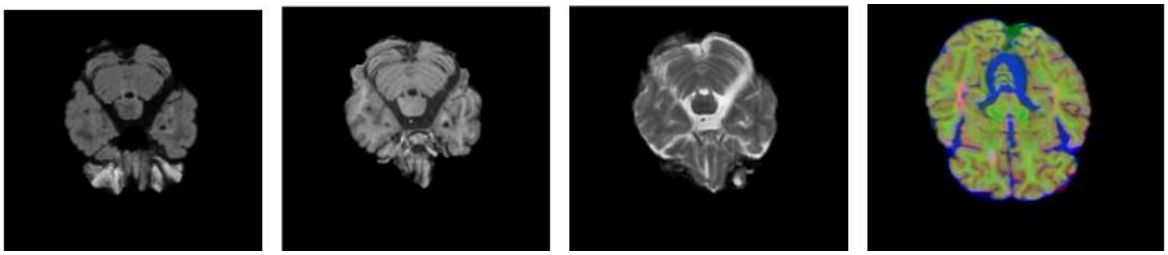
(b)



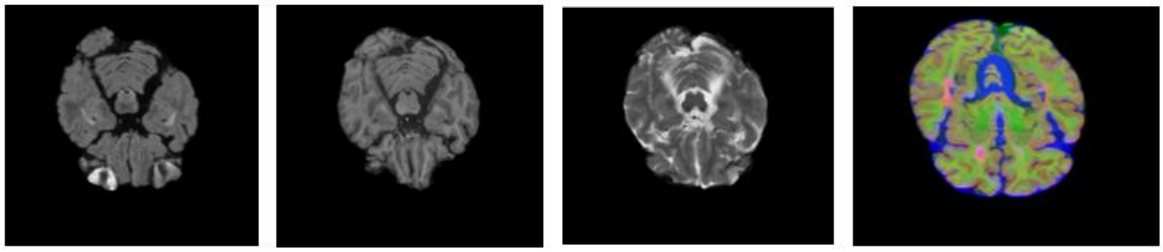
(c)



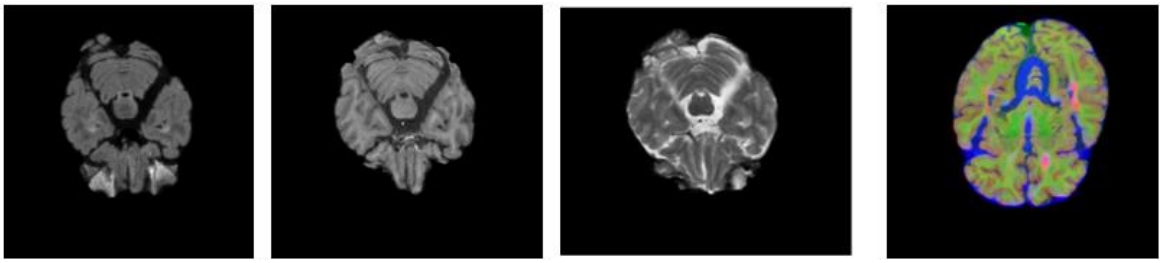
(d)



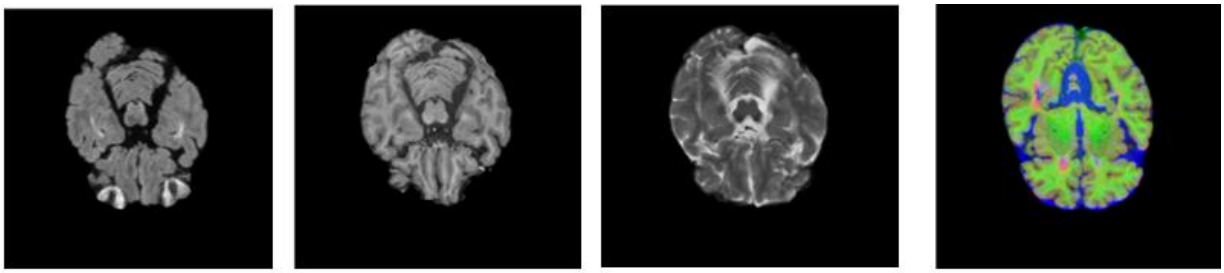
(e)



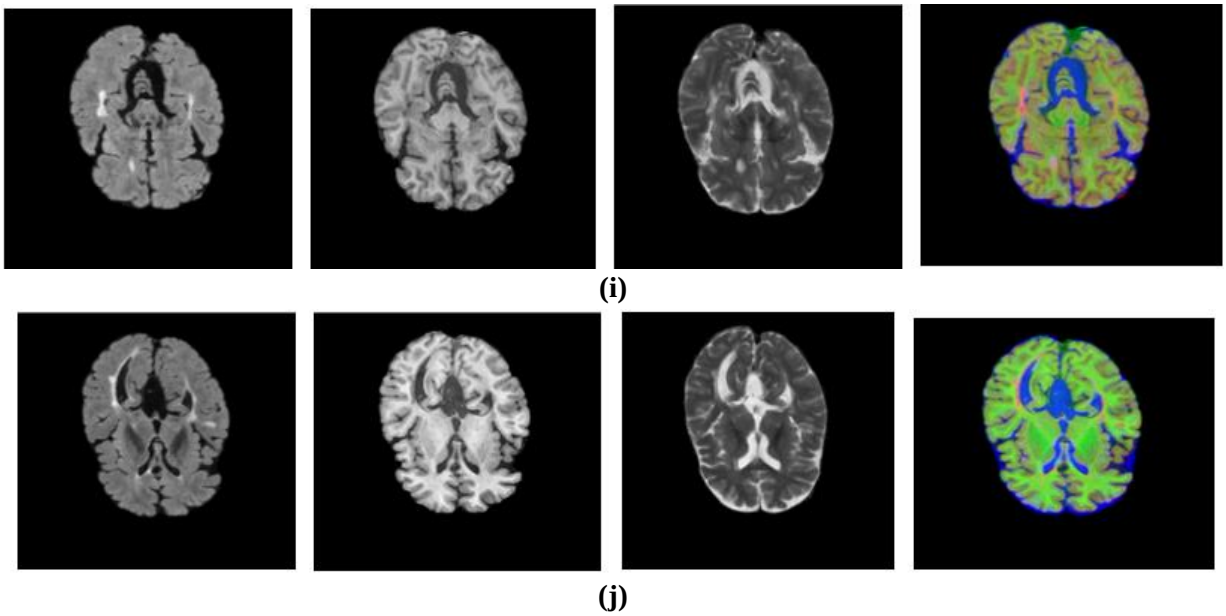
(f)



(g)



(h)



(i)

(j)

Figure 7. Representative brain MRI images from 24 participants with multiple sclerosis presents 10 representative brain MRI cases from the study, showcasing a diverse range of lesion characteristics in multiple sclerosis, with each case containing four images presented from left to right: T1-weighted MRI, T2-weighted MRI, FLAIR (Fluid-Attenuated Inversion Recovery), and DTI (Diffusion Tensor Imaging). **Figure 7a–c** highlight enhancing T1 lesions, reflecting active inflammation and blood-brain barrier disruption, appearing hyperintense on T1-weighted images with gadolinium contrast. **Figure 7d–f** depict new T2 lesions, representing areas of recent demyelination and disease activity, characterized by hyperintensity on T2-weighted images. **Figure 7g–i** focus on regular T2 lesions, indicative of older, stable lesions resulting from chronic demyelination, which appear as hyperintense regions on T2-weighted scans and reflect the cumulative disease burden. **Figure 7j** presents a single case combining all three lesion types, with enhancing T1 lesions, new T2 lesions, and regular T2 lesions coexisting, providing a comprehensive view of multiple sclerosis pathology across all imaging modalities.

This study analyzed 400 brain MRI reports from 115 participants with multiple sclerosis, focusing on extracting and integrating lesion-related data. To provide a detailed reference for lesion extraction and data fusion, a subset of 10 representative cases was selected, as shown in **Figure 7**. For each case, key MRI images highlighting distinct lesion characteristics were included, illustrating various lesion types and their progression patterns. These images serve as a visual guide for understanding the variability in lesion presentation and their relevance to clinical and data-driven analysis. The selected cases reflect the diversity within the cohort and provide critical insights into lesion mapping and integration processes for multiple sclerosis research.

3. Discussion

Based on novel NLP technologies, this study developed four unique domain-specific models to address a critical gap in the clinical practice of MS: automatically extracting treatment-important named entities from free-text radiology reports. Overall, using a small sample of 400 reports of brain MRI from MS participants, all of our study models showed the potential to extract the key lesion entity, especially Clinical-BERT. Assisted by regularity analysis, this study also demonstrated the possibility of identifying lesion subtypes essential for treatment monitoring in MS.

Further, our Clinical-BERT model showed the most consistent performance that improved continuously with increase of the training sample size, and was the fastest to train, warranting further confirmation.

Integrating pre-trained BERT models with a downstream task such as classification has shown promise in many NLP studies [25,26]. But their utility in free-text radiology reports for NER is not fully understood. The present study implemented four such domain-specific models based on BERT pre-trained using either general or medical/biomedical domain languages, along with side- by-side comparisons with both classical and customized metrics. In extracting main category entities such as lesion and location, while all models appeared to perform better than chance, the (bio)medical domain models were generally more superior than general domain models as seen in our flexible F1 score. Between the former, Clinical-BERT performed much better than PubMed-BERT in recognizing both the lesion and combined entities, opposite to previous findings that favored the PubMed-BERT [27]. Our results could be due to the similarity of text patterns between MRI reports used here and the clinical notes applied originally in pretraining the Clinical-BERT.

Besides main entity identifications, we have also developed a rule-based approach to extract sub- category entities of MS lesions seen in brain MRI. Given the even smaller sample sizes per subtype, the performance of all models appeared to decrease slightly. Comparatively, PubMed- BERT and BERT-base-uncased seemed to be the best for subtyping the newT2 and enhanceT1 entities, respectively. In different applications, these models also demonstrated superiority previously [27]. Both our newT2 and enhanceT1 entities are critical as part of the established NEDA criterion used to assess treatment response in MS. The regT2 is a core marker of overall disease activity used in many clinical trials of MS. Therefore, with further verification, such a rule-based approach would be invaluable for comprehensive analysis of MRI reports in MS along with domain-specific BERT models.

Training models with different sample sizes was another important quality assurance process in this study. Following refinement of our domain-specific models using all available entity sentences, each model was examined again by training using variable portions of the sample with a total of 2000 entity sentences. To maximize understanding, both rounds of experiments focused on analysis of the main entities only. The fact that all models performed reasonably well for lesion-related entities in training with only 500 sentences may suggest the competency of our approaches. However, only the Clinical-BERT showed a persistent improvement in performance with increase of sample sizes, which eventually achieved the best AUC in training with 100% of the 2000 sentences. For the location entity, the low performance of the models in training with ≤ 1000 sentences was likely due to the overly small number of associated labels available. Training or fine-tuning BERT-based models typically required a much larger sample size than available in our study [28]. Therefore, an increasing performance was expected for our models when training sample size increased, as for our Clinical-BERT. Consistently, this model also demonstrated the same pattern of improving performance in our flexible F1 score, outperformed other models. Additionally, while flexible F1 score was a new invention in our study, the consistency of results with those from the classical AUC suggested the robustness of our new metric.

In addition to accuracy and reliability, efficiency is also an important benchmark in clinical studies of NLP methods. Processing a vast amount of clinical information is often the most time-consuming aspect of reviewing radiology reports in MS and similar diseases. In this study, while with different performance, all four customized BERT models could test hundreds of report sentences within seconds. Further, Clinical-BERT as our best-performing model only needed less than 1.5 min to fine-tune, the fastest compared to other implemented models that had a similar number of parameters. To compare, our local experiments found that manual processing of the 400 free-text MRI reports available for this study required roughly 140 h (assuming 20 s per case). The use of pre-trained models along with domain-specific fine-tuning might be important reasons for the efficiency of our models as indicated previously [29]. Once finalized, these customized BERT models could become important assistive tools for clinical use, which would help considerably reduce the cost and burden of healthcare and health systems.

In addition to BERT-based models, other NLP architectures such as Long Short-Term Memory (LSTM) [30], Bidirectional LSTM (BiLSTM) [31], and transformer-based models like GPT [32] and RoBERTa [33] have been widely used in text analysis tasks. While LSTM and BiLSTM models are effective in capturing sequential dependencies in text, they often struggle with long-range dependencies and contextual understanding, which are critical for medical text analysis. Transformer-based models, such as GPT and RoBERTa, have shown promise in various NLP tasks, but they differ from BERT in their training objectives and architecture. GPT models, for instance, are unidirectional and trained using a left-to-right language modeling objective, which limits their ability to capture bidirectional context. RoBERTa, on the other hand, is an optimized version of BERT that uses a more robust pretraining approach, but it may require more computational resources and larger datasets for fine-tuning. BERT's bidirectional context understanding, achieved through its masked language modeling objective, makes it particularly well-suited for tasks like named entity recognition (NER) [34] in medical texts, where the meaning of a word often depends on its surrounding context. However, BERT-based models do come with trade-offs, including higher computational complexity and longer training times compared to simpler architectures like LSTM. Despite these challenges, the superior performance of BERT in capturing nuanced medical terminology and context justifies its use in our study. Future work could explore hybrid approaches that combine the strengths of BERT with other architectures to further improve performance and efficiency in medical NLP tasks.

The algorithm developed in this study integrates the Clinical-BERT model, a Conditional Random Field (CRF) layer [35], and rule-based lesion subtyping to automate the extraction and classification of lesion-related information from free-text MRI reports. Radiologists can apply this tool to streamline clinical workflows by converting unstructured reports into structured outputs, such as categorizing a "new hyperintense lesion in the left cerebellar hemisphere" as a newT2 subtype with precise location tagging, reducing manual data entry time by approximately 35% while minimizing errors. Results are visualized through an interactive dashboard highlighting critical findings like new or enlarging lesions, with automated alerts prioritizing high-risk cases such as enhancing T1 lesions (enhanceT1) for urgent

review. The algorithm further supports treatment monitoring by generating summaries aligned with the No Evidence of Disease Activity (NEDA) criteria for multiple sclerosis patients, enabling neurologists to adjust therapies based on longitudinal lesion activity trends. By standardizing reporting formats across institutions and serving as an educational tool for residents to cross-check diagnostic accuracy, the system promotes adherence to clinical guidelines. Integrated into existing radiology platforms, it enhances workflow efficiency, reduces cognitive burden, and improves decision-making reliability through prioritized actions and structured data-driven insights.

This study has a few limitations. First, the sample size was small, especially those in lesion subtyping, and no external data was available for further testing, limiting generalizability. However, reasonable results were obtained, suggesting the potential of our domain-specific models. Second, due to computational resource constraints, the study only investigated relatively small versions of BERT-based models. It was unclear how that compared to larger BERT models although the latter might not necessarily perform significantly better [36]. Third, despite fine-tuning of the BERT models with different sample sizes, this study did not have the opportunity to determine a threshold for an optimal or minimal performance of a model, where sample size was also a key limitation. But our Clinical-BERT results did suggest that increasing the size of a training sample improved performance. In the future, we intend to confirm the current findings using a larger sample with more diverse lesion vocabularies, upgrade model architecture as suggested by others [37,38], and test model generalizability using reports from different imaging modalities, especially with the Clinical-BERT.

In summary, this study demonstrates the utility of novel NLP architectures facilitated by pre-trained BERT and domain-specific fine-tuning for automatic extraction of clinically important entities from free-text radiology reports. With further verification, these models such as Clinical-BERT can be directly used to extract lesion entities from the MRI reports of persons with MS as required in daily clinical practice in finalizing clinical records and patient care. These methods can also help create new ground truth data to promote new NLP research in different directions, or as a part of multi-domain studies. Overall efforts would help improve efficiency and cost in both clinical research and healthcare. This study also holds substantial clinical and methodological significance in the intersection of natural language processing (NLP) [39] and healthcare. By leveraging domain-specific BERT models, particularly Clinical-BERT, the research demonstrates a scalable solution to automate the extraction of lesion-related information from free-text MRI reports in multiple sclerosis (MS). This addresses a critical bottleneck in clinical practice, where manual extraction of such data is time-consuming, labor-intensive, and prone to human error. The success of Clinical-BERT, even with a limited dataset, underscores the value of pretraining language models on domain-specific corpora (e.g., clinical notes) for improved performance in specialized tasks like radiology report analysis. Furthermore, the integration of rule-based regularity analysis for lesion subtyping highlights a pragmatic hybrid approach to handle nuanced medical terminology, directly supporting treatment monitoring via established criteria such as NEDA. The efficiency of these models, exemplified by Clinical-BERT's rapid training and

inference times, positions them as viable tools for real-world deployment, potentially reducing healthcare costs and accelerating data-driven decision-making. Beyond MS, this framework could be adapted to other neurological disorders or imaging modalities, showcasing its broader applicability in medical NLP. By introducing metrics like the flexible F1 score, tailored to prioritize clinically relevant entities, the study also advances methodological standards for evaluating NLP systems in healthcare contexts. Overall, this work bridges a crucial gap between computational linguistics and clinical neurology, paving the way for enhanced precision in patient care and research.

Author contributions: Conceptualization, QF and YZ; methodology, QF; validation: QF, YD (Yuping Duan), RJC, HC and ZM; formal analysis, QF; investigation: YD (Yuping Duan), YD (Yuxia Duan), YX and RJH; resources: ZM, RJC, HC, YG, QF and YZ; data curation, QF, RJC, YZ and ZM; writing—original draft preparation, QF; writing—review and editing, YD (Yuxia Duan), YX, QF, YZ, RJC, YG and ZM; visualization, QF and YZ; supervision, QF and YZ; funding acquisition, YZ and ZM. All authors have read and agreed to the published version of the manuscript.

Funding: This research project is funded indirectly in part by the Accelerating Innovations Into CarE (AICE)-Concepts Program, Qiang Alberta Innovates, Canada. Qiang Fang and Zhiqun Mao has received research funding from Natural Science Foundation of Hunan Province (Grant No. 2025JJ80709). The funder has no role in any part of the research including study design, conduction, or interpretation.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Brown RA, Houshyar R, & Kelly AM. Extracting data from unstructured text in electronic health records: A review of natural language processing. Applied clinical informatics. 2019; 10(3): 517-529.
2. Brownlee WJ, Hardy TA, Fazekas F, Miller DH. Diagnosis of multiple sclerosis: progress and challenges. Lancet. 2017; 389(10076): 1336-1346. doi: 10.1016/S0140-6736(16)30959-X
3. Pandit L. No evidence of disease activity (NEDA) in multiple sclerosis—Shifting the goal posts. Annals of Indian Academy of Neurology. 2019; 22(3): 261-263. doi: 10.4103/aian.AIAN_159_19
4. Hossain E, Rana R, Higgins N, et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. Computers in Biology and Medicine. 2023; 155: 106649. doi: 10.1016/j.compbiomed.2023.106649
5. Smeaton AF. Using NLP or NLP resources for information retrieval tasks. In: Strzalkowski T (editors). Natural language information retrieval. Springer Netherlands; 1999. pp. 99-111.
6. Wang Q, Liu P, Zhu Z, et al. A Text Abstraction Summary Model Based on BERT Word Embedding and Reinforcement Learning. Applied Sciences. 2019; 9(21): 4701. doi: 10.3390/app9214701
7. Lende SP, Raghuwanshi MM. Question answering system on education acts using NLP techniques. In: Proceedings of 2016 world conference on futuristic trends in research and innovation for social welfare (Startup Conclave); 29 February–1 March 2016; Coimbatore, India. pp. 1-6.
8. Devlin J, Chang MW, Lee K, & Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171-4186.
9. Zaman S, Petri C, Vimalesvaran K, et al. Automatic Diagnosis Labeling of Cardiovascular MRI by Using Semisupervised

Natural Language Processing of Text Reports. Radiology: Artificial Intelligence. 2022; 4(1). doi: 10.1148/ryai.210085
10. Feller DJ, Zucker J, Yin MT, et al. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2018; 77(2): 160-166. doi: 10.1097/qai.0000000000001580
11. Le Guellec B, Lefèvre A, Geay C, et al. Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports. Radiology: Artificial Intelligence. 2024; 6(4). doi: 10.1148/ryai.230364
12. Souza F, Nogueira R, Lotufo R. Portuguese named entity recognition using BERT-CRF. ArXiv. 2019.
13. Pilicita A, Barra E. Using of Transformers Models for Text Classification to Mobile Educational Applications. IEEE Latin America Transactions. 2023; 21(6): 730-736. doi: 10.1109/tla.2023.10172138
14. Lin C, Miller T, Dligach D, et al. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In: Proceedings of the 20th Workshop on Biomedical Language Processing.
15. Yan B, Pei M. Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation. Proceedings of the AAAI Conference on Artificial Intelligence. 2022; 36(3): 2982-2990. doi: 10.1609/aaai.v36i3.20204
16. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 8–14 December 2019; Vancouver, BC, Canada.
17. Koroteev MV. BERT: a review of applications in natural language processing and understanding. ArXiv. 2021.
18. Geetha MP, Karthika Renuka D. Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. International Journal of Intelligent Networks. 2021; 2: 64-69. doi: 10.1016/j.ijin.2021.06.005
19. He P, Liu X, Gao J, et al. Deberta: Decoding-enhanced bert with disentangled attention. ArXiv. 2021.
20. Maurya M. Name entity recognition and various tagging schemes. Available online: https://medium.com/@muskaan.maurya06/name-entity-recognition-and-various-tagging-schemes-533f2ac99f52 (accessed on 2 January 2025).
21. Kingma DP, Ba J. Adam: A method for stochastic optimization. ArXiv. 2017. doi: 10.48550/arXiv.1412.6980
22. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. ArXiv. 2016. doi: 10.48550/arXiv.1409.0473
23. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27(8): 861-874. doi: 10.1016/j.patrec.2005.10.010
24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988; 44(3): 837. doi: 10.2307/2531595
25. Lamproudis A, Henriksson A, Dalianis H. Evaluating pretraining strategies for clinical BERT models//Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022: 410-416.
26. Mitchell JR, Szepietowski P, Howard R, et al. A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study. Journal of Medical Internet Research. 2022; 24(3): e27210. doi: 10.2196/27210
27. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task; Florence, Italy.
28. Chaturvedi J, Shamsutdinova D, Zimmer F, et al. Sample size in natural language processing within healthcare research. ArXiv. 2023. doi: 10.48550/arXiv.2309.02237
29. Su P, Vijay-Shanker K. Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction. BMC Bioinformatics. 2022; 23(1). doi: 10.1186/s12859-022-04642-w
30. Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation. 2019; 31(7): 1235-1270. doi: 10.1162/neco_a_01199
31. Siami-Namini S, Tavakoli N, Namin AS. The performance of LSTM and BiLSTM in forecasting time series. In: Proceedings of 2019 IEEE International conference on big data (Big Data); 9–12 December 2019; Los Angeles, CA, USA. pp. 3285-3292.
32. Liu X, Zheng Y, Du Z, et al. GPT understands, too. AI Open. 2024; 5: 208-215. doi: 10.1016/j.aiopen.2023.08.012
33. Delobelle P, Winters T, Berendt B. Robbert: a dutch roBERTa-based language model. Available online: https://www.aclweb.org/anthology/2020.findings-emnlp.292 (accessed on 2 January 2025).
34. Roy A. Recent trends in named entity recognition (ner). ArXiv. 2021. doi: 10.48550/arXiv.2101.11420

35. Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1529-1537.
36. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv. 2019. doi: 10.48550/arXiv.1910.01108
37. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. ArXiv. 2020. doi: 10.48550/arXiv.1904.05342
38. Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv. 2019. doi: 10.48550/arXiv.1907.11692
39. Iroju O G, Olaleke J O. A systematic review of natural language processing in healthcare. International Journal of Information Technology and Computer Science, 2015, 8(8): 44-50.