Natural language processing to estimate RECIST response in cancer patients

Sara Contu
Epidemiology & Biostatistics Department,
Centre Antoine Lacassagne
Nice, France
sara.contu@nice.unicancer.fr

Renaud Schiappa
Epidemiology & Biostatistics Department,
Centre Antoine Lacassagne
Nice, France
renaud.schiappa@nice.unicancer.fr

Emmanuel Chamorey
Epidemiology & Biostatistics Department,
Centre Antoine Lacassagne
Nice, France
emmanuel.chamorey@nice.unicancer.fr

***Abstract*— Treatment response in oncology trials, evaluated according to the Response Evaluation Criteria in Solid Tumors (RECIST) guidelines, is commonly reported as free text in radiology reports, which require manual review for the response extraction. Moreover, in clinical practice, reports do not explicitly mention the RECIST response. We fine-tuned two deep- learning models for Natural Language Processing (NLP), CamemBERT and DrBERT for response classification from radiology reports and we tested them on less structured reports collected in clinical routine to detect instances of disease progression. Both model performed well on reports collected in trials settings (accuracy > 97%), and CamemBERT classified progression in less structured reports with 90% accuracy. Our study provides evidence on the feasibility of using NLP models to determine the treatment response from clinical notes.**

Keywords—NLP, Cancer, RECIST

I. Introduction

Accurate assessment of treatment efficacy of solid tumors is commonly based on the Response Evaluation Criteria in Solid Tumors (RECIST) [1]. RECIST outcomes have played an increasing role in the regulatory drug approval of oncologic therapies and the current version 1.1 is nowadays the most widely used criteria in oncology clinical trials [2].

The evaluation of treatment response is based on the comparison of tumor size assessed by imaging techniques before the treatment onset and at specific times during the course of the treatment. The response falls into four categories: complete response (CR), partial response (PR), progressive disease (PD), or stable disease (SD) [1]. The detection of PD is particularly important as interventions in oncology are often evaluated in terms of progression-free survival (PFS), defined as the time from initiation of treatment to the detection of progressive disease or to the patient's death.

For clinical studies, trial radiologists are chosen to assure data consistency and RECIST outcomes are well reported. However, manual review of the reports by data entry operators is necessary to abstract the response to treatment from the text, a task that is time-intensive and error-prone. Moreover, in clinical routine clear responses according to RECIST are not always reported and scans of the same patient are often evaluated by several radiologists who report their findings in the form of free text with non-uniform structure and lexique. Considering that more than 90% of cancer patients do not participate in clinical trials [3, 4], PFS data and evidence on treatment efficacy are currently underexamined. In clinical practice, tumor response might be abstracted by manual review of reports, but this highly relevant task, which require a comprehensive knowledge, is too time consuming for radiologists.

Natural language processing (NLP) has been applied on radiology reports for a wide variety of tasks, including cancer detection [5], classification of changes in pulmonary nodules [6], follow up recommendations detection [7]. Transformer-based language models trained on radiology reports have also been proposed in English [8]. In this work, we propose the abstraction of the RECIST response from radiology reports and the detection of occurrence of progression using pretrained models on French radiology reports. We compared the performance of a generic model for NLP of French text, named CamemBERT [9], against a model specifically created for NLP of biomedical text, named DrBERT [10].

II. Methods

A. Data and ethical considerations

A total of 2901 radiology reports on cross sectional imaging were collected as a development set from 549 patients participating in clinical trials between 2015 and 2023 for the treatment of solid cancers.

Reports were anonymized and the results section extracted and used as input for the fine-tuning of the models and test (development phase). The corresponding RECIST outcomes were extracted from the electronic Case Report Forms (eCFR) centralizing the RECIST reads available at the Centre Antoine Lacassagne in Nice, France. The eCRF was filled by 21trained radiologists and monitored to ensure data quality. Among reports, 624 revealed a CR, 440 a PR, 948 a SD and 889 a PD.

In order to analyze the possibility to detect PD from radiology reports collected in clinical routine, 120 reports mentioning a status of PD were anonymized for analysis and formed the external test dataset.

According to French regulations, no written informed consent was required, but all patients were informed of the use of their data.

B. Models and evaluation

In order to predict the RECIST response from radiology notes, we fine-tuned CamemBERT and DrBERT. CamemBERT is a state-of-the-art language model based on the RoBERTa architecture model pre-trained on the non-medical text corpus in French from the OSCAR corpus [9]. DrBERT uses the CamemBERT configuration but was trained on an open source corpus of French medical crawled textual data called NACHOS [10].

The models were fine-tuned using cross-entropy loss and Adam optimization with a learning rate of 2×10^{-5}, as used in [11], and we set the batch size to 10. The development set was split into train and test sets (70:30). We evaluated accuracy,

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00178

precision, recall and F1 scores on the development test set. We also applied the models on the external test set to analyze the accuracy of the classification of PD by the two models after the fine-tuning. For the external test, models were applied on the result section of the reports and on full reports due to differences in structures compared to the training set.

III. Results

The CamemBERT model, fine-tuned on our radiology reports, achieved an average accuracy of 97%, while the DrBERT model achieved 98% on the test set used in the development phase. Results in terms of Precision, Recall and F1 score are shown in Table 1. The prediction was based on the results section of the standardized reports collected during clinical trials.

TABLE I. Results on the development test set. CB refers to CamemBERT while DB refers to DrBERT.

Class	Performances					
	Precision		*Recall*		*F1 Score*	
	CB	***DB***	***CB***	***DB***	***CB***	***DB***
PD	0.94	0.98	0.97	0.97	0.95	0.97
CR	0.97	0.98	0.99	1.00	0.98	0.99
PR	0.98	0.98	0.97	0.99	0.98	0.99
SD	0.98	0.99	0.96	0.98	0.97	0.99

For the external test, models were applied on reports from patients not participating in clinical studies but for whom the radiologist noted a progression in the conclusion section of the report. Results showed a superior performance of CamemBERT compared to DrBERT, with the first model reporting an accuracy of 47% of the report as referring to PR compared to 33%. We found a higher accuracy when the full anonymized report was used for the prediction (90% for CamemBERT and 60% of DrBERT) compared to the prediction on the results section.

IV. Discussions

Radiology reports are a valuable source of information for cancer treatment. However, data are unstructured and in the form of free text. Our study analyses whether NLP can be used to abstract treatment response in the context of clinical trials according to the RECIST guidelines and detect the occurrence of progression in reports collected for clinical routine. We have compared two deep learning models trained on general text (i.e. CamemBERT) and medical text (DrBERT) in French on highly variable reports created by different radiologists.

Both models showed very good performances on the set of reports collected in trial settings. We found that when applied on less structured reports collected during clinical practice, performance dropped. Instances of progressive disease were better detected when the full report was used instead of just the result section, as reports did not follow a standardized structure and different pieces of information were potentially scattered in different parts of the report. It is also possible that both models required more resources to fine-tune to generalize well on reports in the external test set. Indeed, BERT models have difficulty in understanding context that is not explicitly stated in the text, and the descriptions reported by the radiologists in the external test data did not include the final evaluation of disease progression or included terms such as "stability of the target lesions" suggesting a SD. Furthermore, is has been previously reported that NLP models show lower performance when divergent semantic tendencies were present, such as the concomitance of progressive disease and improvement, stable disease and worsening, partial response and worsening [12]. Reports collected in clinical routine were generally longer and contained findings of several organs or regions of the body, along with more impressions that made the task more complex. Despite evidence of better results in biomedical tasks obtained by models pre-trained on medical corpora [13], we found that CamemBERT was better suited for detecting PD. This could be related to its ability to capture the nuances of the French language, as large corpora of text data has enabled it to learn more about the language. Future work includes developing a more generalizable model thanks to an improved quality of training data.

The field of NLP is expected to continue advancing, with new techniques and algorithms that possess better abilities to extract contextual information. It might provide RECIST measurements in clinical practice, reducing the burden of manual verification by radiologists on scans that are often handled by several imaging specialists along the treatment.

References

[1] P. Therasse et al., "New guidelines to evaluate the response to treatment in solid tumors," Journal of the National Cancer Institute, vol. 92, no. 3, pp. 205-216, 2000.

[2] E. A. Eisenhauer et al., "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)," European journal of cancer, vol. 45, no. 2, pp. 228-247, 2009.

[3] J. M. Unger, E. Cook, E. Tai, and A. Bleyer, "The role of clinical trial participation in cancer research: barriers, evidence, and strategies," American Society of Clinical Oncology Educational Book, vol. 36, pp. 185-198, 2016.

[4] J. M. Unger, R. Vaidya, D. L. Hershman, L. M. Minasian, and M. E. Fleury, "Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation," JNCI: Journal of the National Cancer Institute, vol. 111, no. 3, pp. 245-255, 2019.

[5] E. Mahoro and M. A. Akhloufi, "Applying Deep Learning for Breast Cancer Detection in Radiology," Current Oncology, vol. 29, no. 11, pp. 8767-8793, 2022.

[6] J. Yuan, H. Zhu, and A. Tahmasebi, "Classification of pulmonary nodular findings based on characterization of change using radiology reports," AMIA Summits on Translational Science Proceedings, vol. 2019, p. 285, 2019.

[7] E. Carrodeguas, R. Lacson, W. Swanson, and R. Khorasani, "Use of machine learning to identify follow-up recommendations in radiology reports," Journal of the American College of Radiology, vol. 16, no. 3, pp. 336-343, 2019.

[8] A. Yan et al., "RadBERT: Adapting transformer-based language models to radiology," Radiology: Artificial Intelligence, vol. 4, no. 4, p. e210258, 2022.

[9] L. Martin et al., "CamemBERT: a tasty French language model," arXiv preprint arXiv:1911.03894, 2019.

[10] Y. Labrak et al., "Drbert: A robust pre-trained model in french for biomedical and clinical domains," bioRxiv, p. 2023.04. 03.535368, 2023.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[12] M. A. Fink et al., "Deep learning–based assessment of oncologic outcomes from natural language processing of structured radiology reports," Radiology: Artificial Intelligence, vol. 4, no. 5, p. e220055, 2022.

[13] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics,* vol. 36, no. 4, pp. 1234-1240, 2020