Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220424

Extraction of Tumor Response Criteria in Semi-Structured Imaging Report

Valentin POHYER[a], David BAUDOIN[a,b], Laure FOURNIER[a,c,d] and Bastien RANCE[a,b,d,e]
[a] *Hôpital Européen Georges Pompidou, AP-HP, Paris, France*
[b] *INSERM, UMRS 1138, Centre de Recherche des Cordeliers, Université Sorbonne-Paris Cité, Paris, France*
[c] *INSERM, PARCC, Paris, France*
[d] *Université de Paris, Paris, France*
[e] *INRIA, France*

Abstract. In this study, we extracted information from 6,376 french CT scan semi-structured text reports evaluating the cancer treatment response using the RECIST methodology. We evaluated the performance against manual annotation of 100 reports and measured the evolution of the presence of information over time. The results show high performances of the extraction as well as trends.

Keywords. information retrieval, natural language processing, imaging, cancer

1. Introduction

Biomedical research used to rely mostly on structured data, however an important part of biomedical information is kept solely in the form of free text reports. Traditional approaches including the identification of regular expressions (regex) have been successfully used in numerous applications [1]. The Response Evaluation Criteria in Solid Tumors (RECIST 1.1) [2] is a methodology used in clinical research and clinical care to evaluate the efficacy of cancer treatments in solid tumors. The radiologists evaluate the evolution of tumors along three axes: target lesions (for which quantitative measurement are performed), non-target lesions, and the appearance of new lesions. An overall conclusion is computed depending on the combination of criteria of the three axes: complete response, partial response, stable disease, progressive disease or not-assessable.

2. Methods

We queried the clinical data warehouse at the European Hospital Georges Pompidou, located in Paris, France, to identify all semi-structured CT reports between January 1st, 2013 and December 31st, 2021. 6,376 reports were identified. We built a set of 243 regex to identify 20 key items using a trial and error process. We applied the regex on the corpus and analyzed errors. We leveraged Py-Rex, a Python tool developed in-house to extract information using the set of regex (code available at:

https://github.com/Vitalizful/Py-Rex). To evaluate the performance, we annotated manually a random sample of 100 documents for three items (baseline date, evolution of the size of target tumors and global RECIST response) and affected the extracted entities to either true positive, true negative, false positive or false negative. We used precision, recall and F1-score to describe the performance of the extractor (see Table 1). To evaluate the stability and quality of the extraction over the study period, we calculated the ratio of the number of data identified over the number of documents in a given period for every trimester.

3. Results and Discussion

Figure 1 shows the evolution of the ratio over the 8 years of the study. Our system captured increasingly more baseline data overtime (increasing from less than 20% of documents to over 60%). Conversely, the identification and extraction of the evolution of the size of target lesion gradually dropped from 90% to 60% over the period of study. Finally, the global RECIST conclusion remains stable overtime at almost 100% ratio of extraction.

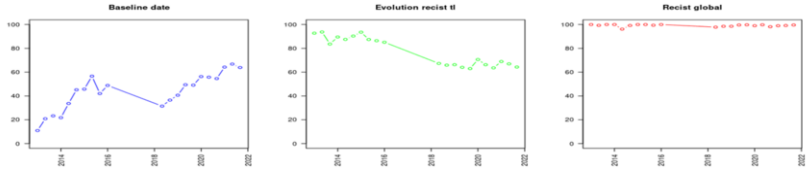

Figure 1. Stability of the extraction over time. The y-axis describes the ratio of documents in which an item is identified.

Table 1. Performance of the extraction

	Precision	Recall	F1-Score
Baseline date	97.8%	95.7%	96.8%
RECIST evolution	97.3%	100%	98.6%
Global conclusion	99%	100%	99.5%

Exchanges between the data scientists and radiologist are required to fully take advantage of the semi-structured documents. The overall performance of our method for the items of interest was very high (over 95%). The application of regex on semi-structured documents considerably simplifies the extraction process. We detected three types of stability profiles: increase, decrease and stability of items. Overall the evolution of the template of the semi-structured documents could be seen as an evolutionary process. The evolution of the structuring of an item is the product of selective pressure, mutation and drift. Our study used solely regex to identify and extract items of interest; however state of the art methods for entity recognition rely on machine learning methods, including neural networks (especially transformers architectures). These methods require large sets of manual expert annotations and often need to be retrained if new elements were to be added. In contrast, regex can be easily developed by radiologists and data scientists together.

References

[1] AAlAbdulsalam AK, et al. Automated Extraction and Classification of Cancer Stage Mentions from Unstructured Text Fields in a Central Cancer Registry. AMIA Jt. Summits Transl. Sci. 2018:16–25.

[2] Eisenhauer EA, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1), Eur. J. Cancer Oxf. Engl. 1990. 45(2009):228–247. doi:10.1016/j.ejca.2008.10.026.