Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210133

Prescreening in Oncology Using Data Sciences: The PreScIOUS Study

Marie ANSOBORLO[a,1], Thibault DHALLUIN[a], Christophe GABORIT[a], Marc CUGGIA[b] and Leslie GRAMMATICO-GUILLON[a]
[a] *CHRU Bretonneau, University Hospital, 2 Boulevard Tonnelé, 37044 Tours, France*
[b] *Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France*

Abstract. The development of precision medicine in oncology to define profiles of patients who could benefit from specific and relevant anti-cancer therapies is essential. An increasing number of specific eligibility criteria are necessary to be eligible to targeted therapies. This study aimed to develop an automated algorithm based on natural language processing to detect patients and tumor characteristics to reduce the time-consuming prescreening for trial inclusions. Hence, 640 anonymized multidisciplinary team meeting (MTM) reports concerning lung cancer were extracted from one teaching hospital data warehouse in France and annotated. To automate the extraction of 52 bioclinical information corresponding to 8 major eligibility criteria, regular expressions were implemented and evaluated. The performance parameters were satisfying: macroaverage F1-score 93%; rates reached 98% for precision and 92% for recall. In MTM, fill rates variabilities among patients and tumors information remained important (from 31.4% to 100%). The least reported characteristics and the most difficult to automatically collect were genetic mutations and rearrangement test results.

Keywords. Lung Neoplasms/statistics and numerical data; Neoplasm Staging/therapeutic use; Multidisciplinary team meeting consultation; Natural language processing; Patient Selection

1. Introduction

The leading cause of mortality by cancer is lung cancer [1]. The burden of lung cancer in France represents more than 30 000 cases and 33 000 deaths in 2018 [2]. With the development of precision medicine, targeted therapies are increasingly studied in clinical trials, especially in lung cancer [3]. Clinical practice guidelines require multiple specific biomarkers testing [4]. To check for clinical trial eligibility, manual review of medical records is essential but a high consuming task in terms of financial and human resources [5]. Increasing difficulties to include new participants into trials is reaching due to numerous and highly specific criteria, potentially cause delay in treatment and opportunity loss for patients [6].

The computerization of the multidisciplinary team meetings (MTM) reports represents a major opportunity to automate the classification of lung cancers to eventually propose trial participation giving the opportunity to receive an innovative therapy. For clinical trial inclusion screening, automatically eligibility criteria checking

[1] Corresponding author, Marie ANSOBORLO, E-mail : Marie.ANSOBORLO@etu.univ-tours.fr.

could reduce the workload and lead to higher efficiency compared to the manual process [7]. Applied in particular in the oncology field, algorithms based on natural language processing (NLP) can be implemented to extract clinical information's for the patient prescreening with very satisfying results [8-10].

Methods based on regular expressions (RegEx) with pattern recognition can extract cancer stage information with high sensitivity [11,12]. Free-text electronic medical reports represent one major information source for NLP methods to identify with high accuracy lung cancer characteristics [10]. Among lung cancer patients, machine learning (ML) method automatically classifying pathology reports for the cancer stage does already exist [8]. Naïve Bayes Classifier methods developed within the Machine Learning community have been successfully used to classify text documents [11]. However, to the best of our knowledge, only few studies have attempted to compare RegEx methods and naïve Bayes Classifier methods to extract information about multiple patients and cancer bio clinical characteristics from MTM records. To increase and simplify patient selection for anti-cancer treatment trials, this study aims to develop and assess two methods based on features extraction and supervised classification for feasibility of PreScreening in Oncology Using data Sciences (PreScIOUS).

2. Methods

The prescreening for bio clinical characteristics was performed on MTM reports from lung cancer patients stored in one hospital data warehouse in France based on the EHOp® model [12]. Complete "thoracic MTM reports" filled between 2018 and 2020 were included and "Non Tracheo Bronchial Tumor" reports were excluded of the study.

The main judgment criterion was the identification of 8 eligibility criteria, representing 52 different levels depending on international nomenclatures. Up to 15 references from the International Classification of Diseases 10th Edition for Oncology (ICD-O 10), were used to define the 4 "Histology" levels. The rarest subtypes were grouped at the "other" level. The "T", "N" and "M" factors were taken from the TNM 2017 classification and the global "TNM stage" has been inferred on the basis of the 8th edition of the TNM by combining the "T", "N" and "M" levels. The World Health Organization Performance status and the ALK gene rearrangement and EGFR receptor mutations were also screened [13,14]. For each factor but WHO PS, missing data in the text corpus has been annotated as an additional modality in the PreScIOUS tool.

A sample of 50% of this MTM were manually annotated to improve algorithms and to constitute the gold standard. MTM sample was split into equal size training and testing sets. The identification performances of these 8 eligibility criteria were precision, recall and F1 score calculated on the results obtained.

The PreScIOUS study was based on three steps: (i) MTM preprocessing to normalize symbols, orthograph and abbreviation on free text, (ii) RegEx and N.B.C. implementation to extract information and classify MTM about eligibility criteria (iii) model performance is evaluated with manual annotation as gold standard. Forty-five distinct regular expression patterns have been built based on the train set and optimized to improve performance compared to manual annotation. Queries on unstructured content were implemented and performed using the R "stringr" package. At the evaluation step, PreScIOUS was once applied on the testing set.

For the machine leaning models a tokenization was performed by converting free text into tabular format using the document term matrix (DTM). DTM values were term frequency-inverse document frequency (TF-IDF) weight a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [15]. Bayesian classifier was optimized to improve the optimal value of the estimator of "Laplacian probabilities" compared to manual annotation. Bayesian classifier were performed from the R package naivebayes (with the naive_bayes function) [16].

3. Results

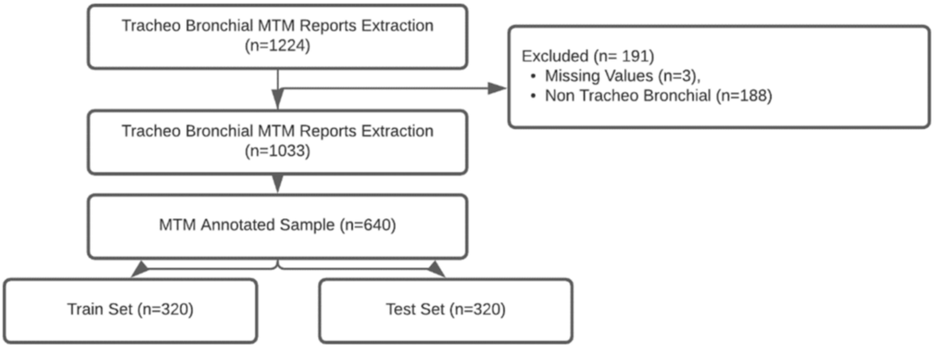
Tracheo Bronchial MTM Reports Extraction
(n=1224)
Excluded (n= 191)
• Missing Values (n=3),
• Non Tracheo Bronchial (n=188)
Tracheo Bronchial MTM Reports Extraction
(n=1033)
MTM Annotated Sample (n=640)
Train Set (n=320)
Test Set (n=320)

Figure 1. Flow chart

A total of 1224 MTM were extracted over the study period, 3 MTMs were not usable, 188 concerned non-tracheal bronchial tumors. Of the 1033 MTM of tracheal bronchial tumors 640 were manually annotated, 320 MTM in the train set 320 MTM in the test set (Figure 1).

PreScIOUS obtained very satisfying precision and recall rates (>80%) for most of the factors. Overall, the RegEx showed better results than the supervised machine learning method using a Naïve Bayes Classifier. The precision rates of RegEx were all above 96% and almost perfect for “WHO PS” (>99.9%). Highest rates were calculated for RegEx with recall over 95% for “T” and “N” and F1-mesures above 98% for “M”, “TNM stage” and “WHO PS” (Table 1).

With the N.B.C. models, most of the factor (62.5%) presented satisfying F1-mesures in macro average (≥74%) (Table 2). Concerning recall, “Histology” and “TNM stage” models reached more than 74%. “EGFR” and “T” models obtained precision rates 85% and 90% and “TNM stage” between 90% and 95%. The lowest F1-mesures were obtained for “EGFR” mutation and “ALK” rearrangement tests results. Few factors as “WHO PS”, “ALK”, “EGFR” presented not satisfying precision or recall rates (< 60%)

The most frequent information in the 320 MTM free texts from the test set were the WHO PS factor with 100% of annotation before histology (>84%). EGFR and ALK were the factors with the highest rate of not retrieved (“nr”) information (>61%) among the MTM free text. Information about T, N or M, was present in the majority of MTM reports (≥61%). Most of the MTM sample (55%) presented information for imputed TNM stage.

The macroaverage F1-mesure differences between the two methods were smaller than 26 points except for the “WHO PS” and “ALK” information. For these 2 factors, the supervised machine learning method failed to correctly classify two levels. These were the two rarest levels “3” and “4” for “WHO PS” and “1” and “nt” for “ALK “. For the level "nt" of "EGFR" factor, the recall rate evaluated by the N.B.C. method was twice higher than the RegEx method.

Table 1. Regular Expression classification performances parameters. Values are percentages.

	Histology	T	N	M	TNM stage	WHO PS	ALK	EGFR
Precision	99.4	99.6	99.3	98.1	96.4	97.0	99.9	96.4
Recall	98.1	95.6	99.2	95.6	75.6	75.9	98.6	94.1
F1	98.8	97.4	99.3	96.8	80.3	78.9	99.2	94.8
Missing	15.3	37.5	37.5	38.4	45.0	0	61.6	65.9

Table 2. Bayesian classifier performances parameters. Values are percentages.

	Histology	T	N	M	TNM stage	WHO PS	ALK	EGFR
Precision	86.6	87.0	76.9	78.7	82.5	56.5	42.3	84.7
Recall	72.3	57.2	72.1	67.9	68.3	42.5	59.6	74.2
F1	77.7	71.8	73.6	71.6	73.6	56.6	41.2	77.8
Missing	15.3	37.5	37.5	38.4	45.0	0	61.6	65.9

4. Discussion

PreScIOUS eases prescreening task with high performances. The main strengths of the study are the ability to classify medical reports for multiple patient and cancer profiling factors with very satisfying performances higher than previous methods [8,10]. This study demonstrates that supervised machine learning models could represents innovative supports for rule-based systems.

Only one reviewer participated in manually annotating the text, representing one methodological weakness of the study. Heterogeneous performances among same factors for distinct labels were observed as consequence as rare levels and unstructured information in MTM report. Among the levels researched globally in the MTM reports sample, few were mostly absent. Information about EGFR results were more frequently reported than for ALK, such as described in literature [14]. For the "ALK" and "EGFR" factors, the precision rate was excellent but the sensitivity to retrieve the information was smaller because of very rare occurrence of "nt" (i.e., "not tested"). As the data of the test set were never used to train the M.L. model, a category that was absent in the train test could not be assigned by the model. Two levels presented very rare observations (i.e., "tis" and "Ia3") and were considered as outliers and recoded as "no reference" ("nr") for N.B.C. model.

Training the model on a larger MTM reports sample could increase the performance and the reproducibility of the machine learning approach to recognize rarest events. Requiring only few data, RegEx are efficiently implemented with help of medical expertise and could be combined to machine learning methods to ease prescreening in oncology on large datasets. The human readable aspect of RegEx allows better results explanation than 'black box' methods such as machine learning.

As MTM reports structure may vary between healthcare settings, the construction of RegEx needs to be improved to be generalizable across multiple facilities. The two PreScIOUS methods merit to be applied in complement on hospital data warehouse from other centers to value their external validity.

References

[1] Ferlay J, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359-386. doi:10.1002/ijc.29210.

[2] Defossez G, et al. Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018 : étude à partir des registres des cancers du réseau Francim. Résultats préliminaires. Synthèse., (n.d.). https://www.santepubliquefrance.fr/docs/estimations-nationales-de-l-incidence-et-de-la-mortalite-par-cancer-en-france-metropolitaine-entre-1990-et-2018-etude-a-partir-des-registres-des (accessed December 13, 2020).

[3] Hirsch BR, et al. Characteristics of Oncology Clinical Trials: Insights From a Systematic Analysis of ClinicalTrials.gov. JAMA Intern Med. 2013;173:972–979. doi:10.1001/jamainternmed.2013.627.

[4] Planchard D, et al. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2018;29: iv192–iv237. doi:10.1093/annonc/mdy275.

[5] Jouis V., Le screening et l'inclusion, 2017: 37.

[6] Unger JM, Cook E, Tai E, and Bleyer A. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. Am Soc Clin Oncol Educ Book. 2016;35:185–198. doi:10.1200/EDBK_156686.

[7] Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, Li Q, Zhai H, and Solti I. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc. 2015;22:166–178. doi:10.1136/amiajnl-2014-002887.

[8] AAlAbdulsalam AK, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM. Automated Extraction and Classification of Cancer Stage Mentions fromUnstructured Text Fields in a Central Cancer Registry. AMIA Jt Summits Transl Sci Proc. 2018;2017:16–25.

[9] Nguyen AN, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc. 2010;17: 440–445. doi:10.1136/jamia.2010.003707.

[10] Mitchell T. Machine learning. 1997.

[11] Madec J, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Studies in Health Technology and Informatics. 2019;264:1536–1537. doi:10.3233/SHTI190522

[12] Bethune G, Bethune D, Ridgway N, Xu Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. J Thorac Dis. 2010; 2: 48–51.

[13] Hofman P. ALK in Non-Small Cell Lung Cancer (NSCLC) Pathobiology, Epidemiology, Detection from Tumor Tissue and Algorithm Diagnosis in a Daily Practice, Cancers (Basel). 9 (2017). doi:10.3390/cancers9080107.

[14] Manning CD et al., Introduction to Information Retrieval, (n.d.). https://nlp.stanford.edu/IR-book/ (accessed December 13, 2020).

[15] Majka M. High Performance Implementation of the Naive Bayes Algorithm, (n.d.). https://majkamichal.github.io/naivebayes/ (accessed November 19, 2020).

[16] McCowan I, Moore D, and Fry M. Classification of Cancer Stage from Free-text Histology Reports, n.d. doi:10.1109/IEMBS.2006.259563