World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:17, No:03, 2023

Clinical Validation of an Automated Natural Language Processing Algorithm for Finding COVID-19 Symptoms and Complications in Patient Notes

Authors : Karolina Wieczorek, Sophie Wiliams

Abstract : Introduction: Patient data is often collected in Electronic Health Record Systems (EHR) for purposes such as providing care as well as reporting data. This information can be re-used to validate data models in clinical trials or in epidemiological studies. Manual validation of automated tools is vital to pick up errors in processing and to provide confidence in the output. Mentioning a disease in a discharge letter does not necessarily mean that a patient suffers from this disease. Many of them discuss a diagnostic process, different tests, or discuss whether a patient has a certain disease. The COVID-19 dataset in this study used natural language processing (NLP), an automated algorithm which extracts information related to COVID-19 symptoms, complications, and medications prescribed within the hospital. Free-text patient clinical patient notes are rich sources of information which contain patient data not captured in a structured form, hence the use of named entity recognition (NER) to capture additional information. Methods: Patient data (discharge summary letters) were exported and screened by an algorithm to pick up relevant terms related to COVID-19. Manual validation of automated tools is vital to pick up errors in processing and to provide confidence in the output. A list of 124 Systematized Nomenclature of Medicine (SNOMED) Clinical Terms has been provided in Excel with corresponding IDs. Two independent medical student researchers were provided with a dictionary of SNOMED list of terms to refer to when screening the notes. They worked on two separate datasets called "A" and "B", respectively. Notes were screened to check if the correct term had been picked-up by the algorithm to ensure that negated terms were not picked up. Results: Its implementation in the hospital began on March 31, 2020, and the first EHR-derived extract was generated for use in an audit study on June 04, 2020. The dataset has contributed to large, priority clinical trials (including International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC) by bulk upload to REDcap research databases) and local research and audit studies. Successful sharing of EHR-extracted datasets requires communicating the provenance and quality, including completeness and accuracy of this data. The results of the validation of the algorithm were the following: precision (0.907), recall (0.416), and F-score test (0.570). Percentage enhancement with NLP extracted terms compared to regular data extraction alone was low (0.3%) for relatively well-documented data such as previous medical history but higher (16.6%, 29.53%, 30.3%, 45.1%) for complications, presenting illness, chronic procedures, acute procedures respectively. Conclusions: This automated NLP algorithm is shown to be useful in facilitating patient data analysis and has the potential to be used in more large-scale clinical trials to assess potential study exclusion criteria for participants in the development of vaccines.

Keywords : automated, algorithm, NLP, COVID-19
Conference Title : ICBB 2023 : International Conference on Bioinformatics and Biomedicine
Conference Location : London, United Kingdom
Conference Dates : March 16-17, 2023