Natural Language Processing and Machine Learning in Electronic Health Records

Ashima Kumar

University of San Diego

MSNC 507: Statistics

Dr. Thidarat Tinnakornsrisuphap

June 26, 2023

Natural Language Processing and Machine Learning in Electronic Health Records

Dr. Robert Y. Lee is a physician and researcher specializing in pulmonary and sleep medicine. He currently holds medical and master's degrees. Furthermore, he is affiliated with the division of pulmonary, critical care, and sleep medicine in the department of medicine at the University of Washington (UW) in Seattle. Dr. Lee further contributes his knowledge through his research and clinical practices.

In this article, the question of whether natural language processing (NLP) is effective in clinical trial outcomes is measured by various methods. NLP is part of the field of artificial intelligence that focuses on the algorithms and syntax between computers and the human language. Its purpose is to generate and interpret human language in a meaningful way. The value of the journal article is addressed by evaluating outcomes of NLP from free-text electronic health records (EHRs), which can be tedious to look through manually by individuals (Lee et al., 2021). This particular research is important for researchers and clinicians alike because it provides perception on the accuracy and efficiency of using NLP in clinical settings. Furthermore, this respective study contributes to the benefits and limitations of using NLP in measuring clinical documented outcomes and can help make better decisions regarding EHR documented goals-of-care.

How the Research was Conducted

The research conducted in this study used the approaches of NLP and machine learning (ML) to identify goals-of-care discussions for patients with serious illnesses in EHRs. The study applied inpatient and outpatient clinical notes from two teaching hospitals and affiliated clinics of UW Medicine (Lee et al., 2021). The data sources were collected from two different randomized control trials of palliative care interventions, a convenience sample written by palliative care specialists that included keywords of goals-of- care, and clinical notes from a random sample of patients with serious illness (Lee et al., 2021). These clinical notes were collected as plain text from the UW Medicine clinical EHR platforms and the objective was to identify goals-of-care discussions.

To define what goals-of-care meant, the researchers in the study developed an operational definition that included advance care planning activities, completion of advance directives of physician order for life-sustaining treatment forms, and referral to hospice or subspecialty palliative care services (Lee et al., 2021). The 1,335 notes that were extracted were labeled based on the presence or absence of goals-of-care discussion using a structured data collection instrument or by study investigators (Lee et al., 2021). Furthermore, NLP and ML methods were used to tokenize the notes into one-word length tokens. The data set was split into 100 pairs of training sets that were assessed using an ML classifier called the scikit-learn library, which was trained to compute a predicted probability of goals-of-care documentation (Lee et al., 2021). In addition, the sample set was separated into inpatient and outpatient sets to assess performance characteristics of NLP/ML performance (Lee et al., 2021). Programming was completed in Python v3.6, natural language processing was used with the Natural Language Toolkit v3.2.4, and ML was conducted by the scikit-learn library v0.19.0 (Lee et al., 2021).

Collected Data, Analysis, and Results

The study used a sample of inpatient and outpatient notes and the application of NLP/ML techniques to extract goals-of-care discussion from EHR notes. The collected data and analysis have shown that 689 of the total 3,183 notes in the data set were identified as goals-of-care positive notes (Lee et al., 2021). Among this data, 265 notes were identified from randomized controlled trials and 424 were identified through manual review. As shown in Figure 2, The sample set was divided into 1,435 inpatient notes which resulted in 468 goals-of-care discussions, and 1,748 outpatient notes in which 221 included goals-of-care discussions (Lee et al., 2021).

The performance of the NLP/ML methods was evaluated using sensitivity, which ranged from 72% to 93%, with a mean sensitivity of 82.3%. As shown in Figure 1, The observed specificity of NLP/ML that identified goals-of-care in the EHR notes ranged from 95% to 99% with a mean specificity of 97.4% (Lee et al., 2021). In Figure 3, the positive likelihood ratio was a median of 32.2 and the negative likelihood ratio was a median of 0.18 (Lee et al., 2021). In addition, on Figure 3, the area under the curve ranged from 0.91 to 0.98 (Lee et al., 2021). The results shown in Figure 1 further showcased that NLP/ML classifiers had better test performance on inpatient records (mean sensitivity 97.1%) compared to outpatient records (mean sensitivity 96.8%) (Lee et al., 2021).

The use of NLP/ML for identifying goals-of-care in EHR notes across inpatient, outpatient and palliative contexts has demonstrated its capability of being useful. Being able to use NLP/ML mechanisms to comb through EHR clinical notes can further improve and help cut down time for clinicians to search for specific sets of data. The collected data from this study emphasize the range of complexity that NLP/ML mechanisms can extract from complex documentation.

Figure 1

Performance Characteristics of NLP/ML by Note Type

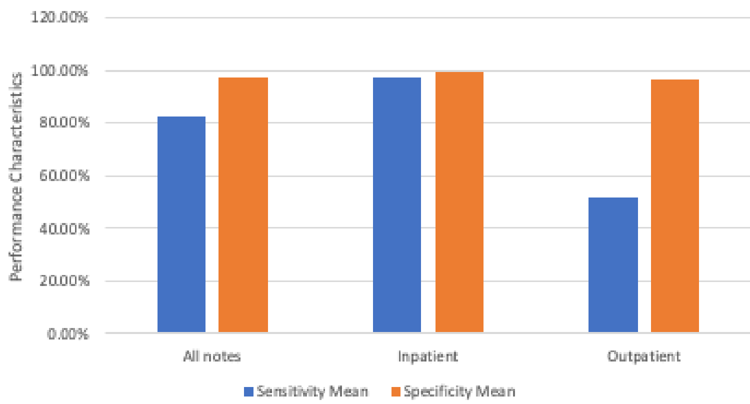


Figure 2

Composition of Sample Dataset

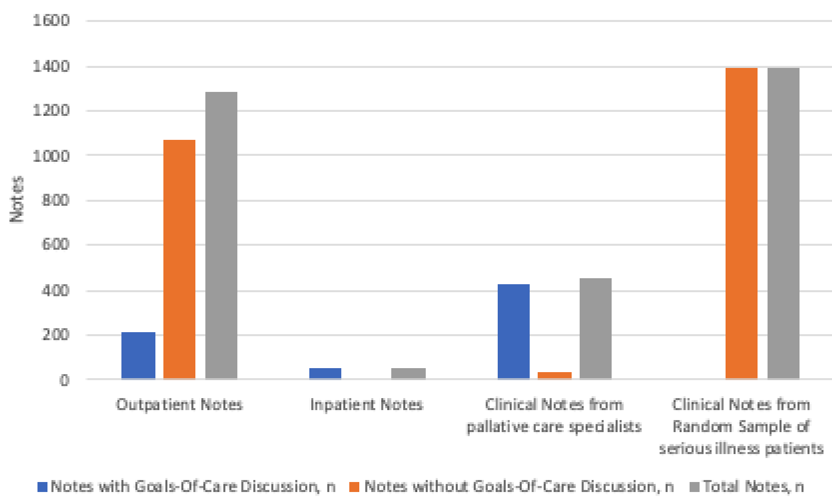
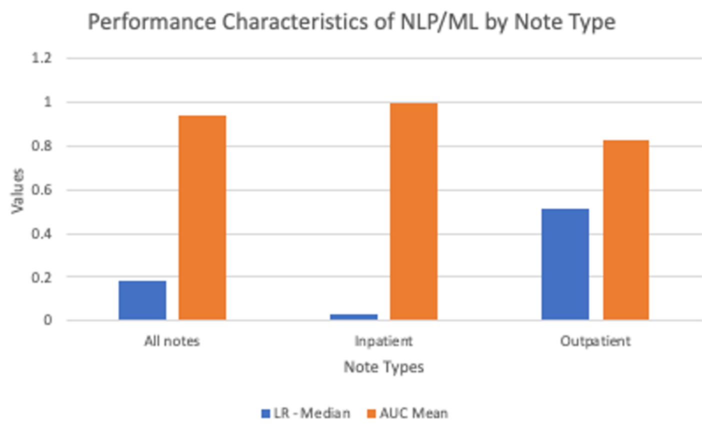


Figure 3

Mean and Median Classification of Note Type



Conclusions

The parameter of interest in this study is the identification of goals-of-care discussions in clinical EHR notes (Lee et al., 2021). Researchers in this study sought to develop a search tool that used NLP and ML techniques to identify keywords goals-of-care in patient notes and discussions (Lee et al., 2021). From the study, 689 goals-of-care discussions were found from inpatient and outpatient notes using NLP and ML methods (Lee et al., 2021). The research question of interest was whether NLP/ML could be effectively used to identify goals-of-care discussions in EHR notes across different clinical settings.

To address these questions, the researchers applied NLP/ML techniques through an automated program named scikit-learn library v0.19.0 to identify goals-of-care in outpatient and inpatient notes (Lee et al., 2021). One of the statistics that was evaluated was sensitivity, which is a measure of how well the program identified goals-of-care documentation through true positives over the total amount of true positives and false negatives. Researchers found a higher average sensitivity of 97.1% on inpatient records compared to outpatient records, which had an average sensitivity of 96.8% (Lee et al., 2021). This statistic displayed an accurate presentation of goals-of-care in the clinical notes. Furthermore, another statistic that played a major role in classifying the efficiency of NLP/ML techniques was specificity which is the probability of clinical notes that did not have goals-of-care in their documentation. An additional statistic that was used was positive and negative likelihood ratios, where the value of the positive likelihood ratio (LR+) measures the change of odds of having goals-of-care discussions from the NLP/ML classifier (Lee et al., 2021).

The findings of this study suggest NLP/ML classifiers can help be used to identify goals-of-care discussions in EHR notes across inpatient and outpatient settings (Lee et al., 2021). These findings indicated high sensitivity and specificity and can be beneficial to the field of palliative care as well as clinical settings across the board. However, one can infer based on the results of the study, inpatient settings have a higher sensitivity due to more detailed and critical notes. Another reason could also be that inpatient settings may have more time-sensitive decisions where the topic of goals-of-care could be discussed thoroughly.

Strengths and Weaknesses of the Selected Statistical Methods

To look at a deeper analysis of the statistics, the data that was collected was not random and was taken from a specific sample of EHRs from UW Medicine. This could be a potential limitation because the data may not accurately reflect how NLP/ML mechanisms could process EHRs across different hospitals and scenarios. The data could be more widespread if there was access to different types of EHRs which could have varying sensitivities and specificities. Furthermore, although the sensitivity and specificity were different values between inpatient and outpatient EHRs, these results could have been specific to this particular facility only. The definition of goals-of-care could also be varied in definition between other facilities which could make NLP/ML processing more difficult. The study’s results are trustworthy because the sample that was used was controlled clinical notes from inpatients and outpatients with serious illnesses which are multiple diverse sources. I believe the diversity of inpatient and outpatient clinical notes enhances the diversity of the findings; however, these notes could be further looked at in different healthcare departments.

Another limitation I would consider in this study is the accuracy of the clinician notes and if goals-of-care discussions on the EHR notes were precise. There is a potential for NLP/ML mechanisms to misclassify goals-of-care in EHRs because the data retrieved only captured sensitivity and specificity, but not the depth of goals-of-care. If the study was able to also measure the quality of the clinical notes by incorporating additional NLP techniques, then I believe the study will be more accurate. In addition, it would be beneficial to focus on different types of NLP/ML mechanisms and compare the average performance of accuracy to each other through a paired *t*-test. If the study could be improved, I believe it would benefit from a larger dataset from different facilities including multiple healthcare systems from different states as well as a wider population of patients. In conclusion, NLP/ML mechanisms used in this study delivered significant results; however, further improvements can be implemented if the study were to be repeated for more accurate results.

References

Lee, R. Y.-F., Lober, W. B., Sibley, J., Kross, E. K., Engelberg, R. A., & Curtis, J. R. (2019). Identifying goals-of-care conversations in the electronic health record using machine learning and natural language processing. *Journal of Pain and Symptom Management*, *61*(1), 136–142. https://doi.org/10.1016/j.jpainsymman.2020.08.024