MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220156

Review of the Performance Metrics for Natural Language Systems for Clinical Trials Matching

Jeongeun Kim[a,b], Yuri Quintana[a,b]

[a]*Harvard Medical School, Boston, MA*
[b]*Beth Israel Deaconess Medical Center, Department of Clinical Informatics, Boston, MA*

Abstract

Natural Language Processing (NLP) has been adopted widely in clinical trial matching for its ability to process unstructured text that is often found in electronic health records. Despite the rise in the new tools that use NLP to match patients to eligible clinical trials, the comparison of these tools is difficult due to the lack of consistency in how these tools are evaluated. The ground truth or reference that the tools use to assess results varies, making it difficult to compare the robustness of the tools against each other. This paper alarms the lack of definition and consistency of ground truth data used to evaluate such tools and suggests two ways to define a gold standard for the ground truth in small and large-scale studies.

Keywords

Natural Language Processing, Clinical Trial Matching, Eligibility Criteria

Introduction

Clinical trials offer an opportunity for better treatments to current patients or future patients and advance scientific research.[1,2,3] A successful clinical trial necessitates accurate and robust cohort selection under specific criteria.[4] The cohort selection can be improved by algorithms that match clinical trials eligibility criteria to data in the electronic health records (EHR).[4] However, EHR itself does not provide sufficient information, as some of the patient information is formatted as unstructured free-text.[4] In addition, the eligibility criteria of the clinical trials are also documented as unstructured free-text.[5] Natural Language Processing (NLP) has been identified as a solution to analyze unstructured text and generate structured representations.[6] NLP uses ground truth to train, test, and evaluate the model and human annotation is deemed the most reliable way of collecting it.[7] In clinical NLP, expert knowledge is expected from the annotators, which can be challenging to find, time-consuming and financially costly.[7]

When natural language processing systems are evaluated, performance metrics such as recall (or sensitivity), precision (or positive predictive value), F-measure (harmonic mean of recall and precision), accuracy, and specificity are commonly reported.[6] However, a 2018 study that compared the concordance of an Artificial Intelligence (AI) clinical decision-support system (CDSS), Watson for Oncology (WFO), and a multidisciplinary tumor board for breast cancer raised the issue of the absence of a gold standard beyond expert opinion which can vary profoundly during the evaluation of CDSS.[8] There could be inherent ambiguity in the text[7] or multiple valid interpretations, thus causing disagreements between reviewers, such as how patients with comorbidities and increasing age are treated.[8]

The discrepancy and unreliability of gold standard or ground truth decrease the performance metrics' validity and make cross-comparison of multiple NLP tools difficult. In this study, the evaluation methods of clinical trial matching systems are investigated for their definitions of gold-standard. This study urges the need for a gold standard in the patient eligibility of clinical trials and encourages studies to report the agreement as a part of their performance metric.

Methods

In May 2021, PubMed, ACM Digital Library, and IEEE Xplore were searched using the following query: clinical trial matching AND natural language processing AND evaluation. ScienceDirect was searched using the following query: clinical trial matching AND natural language processing. The queries were used to search against all fields in PubMed and title only in the other three databases. Only the articles published between 2016 and 2021 were considered. Papers were filtered if they did not directly evaluate the performance of clinical trial matching systems. Articles were reviewed for four main pieces of information: 1) clinical trial matching system used, 2) performance reported, 3) the definition of their gold standard, 4) any reports on the agreement among the reviewers for inter- and intra-rater reliability. If the gold standard was defined in the article, the reviewers' expertise (clinician or non-clinician) was also noted.

Results

Table 1 shows the articles for the query clinical trial matching AND natural language processing AND evaluation published between 2016 - 2021. Five articles were evaluated for the performance of clinical trial matching models. 2 articles used the IBM Watson for Clinical Trial Matching, 1 article used concept2vec, and 2 articles used unnamed models, one of which was a machine learning classifier with named entity recognition and the other was a SVM classification method.

All five articles reported their gold standard or standard reference definitions, but with varying methods. Of the five articles, four articles conducted a manual review by multiple raters. However,

not all studies that used multiple raters reported Cohen's kappa. Furthermore, the number and the medical background of these raters were diverse and inconsistent; clinicians, annotators, nurse coordinators, oncologists, trial team, and trainees of an informatics program. One article had a unique definition of the reference standard, which was the original clinical enrollment status of the patients.

System Name	Performance	Gold Standard	Report agreement?	citation
IBM Watson for clinical trial matching	Agreement Level: >= 97%, accuracy: 91.6%, recall: 83.3%, precision: 76.5%, NPV: 95.7%, specificity: 93.8%	Manual review by two clinicians, with discrepancies, discussed to achieve consensus	Yes	9
concept2vec	F1 score: 0.8038	Annotators analyzed each of the patient records according to 13 criteria to decide about their eligibility	No	10
Watson for Clinical Trial Matching	Agreement: 81%-96% Specificity: 76 - 99%, sensitivities: between 91% and 95% for three trials and 46.7% for the 4th	Manually reviewed by clinical staff (nurse coordinators, then oncologist and trial team)	Yes	11
No Name	F2 score: 0.85, 0.91	Twelve volunteers recruited among National Library of Medicine informatics program trainees	Yes	12
No Name	AUC: 75.5% - 89.8%, Recall: 70.6% - 79%, MAP: 18% - 35.2%	Original clinical enrollment status	No	13

Table 1. Summary of the evaluation methods and clinical trial matching system used in 5 articles that evaluated the performance of clinical trial matching systems between 2016 and 2021.

Discussion

Manual review (four out of five studies) was the most common method of defining the ground truth. However, not all of these studies reported Cohen's kappa, and the expertise and number of the reviewers used varied. Both Alexander *et al*. and Beck *et al*. used IBM Watson for Clinical Trial Matching. In contrast, Alexander *et al*. had two clinicians manually review the patient-trial matches[9], Beck *et al*. did not specify the number of reviewers nor their expertise.[11] Such an approach of obtaining the ground truth from domain experts is costly and susceptible to poor quality.[7] First, recruiting experts to label data and training them is financially costly.[7,15] Second, labeling and annotation is a

highly time intensive process.[7, 15] Third, the process of eliminating a disagreement between reviewers may introduce artificial data that is neither general, nor reflects the ambiguity inherent in natural language due to the use of overly prescriptive annotation guidelines.[7]

Despite the shortcomings of manual review by experts, it is important to understand the role of humans in natural language processing. Therefore, this study suggests two approaches for small and large-scale studies that utilize humans yet are more scalable and cost-efficient. As Meystre *et al*. have defined, small-scale studies should use the patient's original clinical enrollment status as the ground truth. The patient enrollment status is an objective metric, free of reviewers' bias or interpretations on a patient's eligibility. However, this may have been possible due to their small sample size (a cohort of 229 patients with breast cancer) and the small number of trials the study looked at (3 trials). For larger-scale studies, an automated process for generating ground truth will be needed to produce training data scalable to the model's size, considering the cost and the unscalable nature of a manual expert review. It would require human input to develop and streamline the process compared to simply scaling up the manual review.

There was heterogeneity in the medical knowledge across reviewers in the four papers that utilized manual review. The expertise of the manual reviewers ranged from annotators without a complete description of their job title or field, to trainees of an informatics program, nurse coordinators, clinicians, and oncologists. However, recent studies found that crowdsourcing for ground truth produced the quality equivalent to what the domain experts would produce.[7, 15] In a similar context, once experts create a guideline of matching patients to a trial based on similarity scores, an automated algorithm, such as the recurrent neural network, could generate similarity scores, translating into ground truth. The automation will help in obtaining a large training dataset. However, such ground truth generators may be challenging to build for rare diseases where it is difficult to reach a big enough training size to train a model. In such a scenario, using citizen science to create quality annotations is a considerable option.

The nature of clinical narratives as a sublanguage makes it difficult to apply NLP trained on the general text and requires domain-specific development and training.[16] Considering the diversity of electronic health record systems used across hospitals, generating training data and ground truth requires study-level efforts. Therefore, a single definition of ground truth must be used across studies so that the NLP tools developed are comparable.

Moving forward, it is essential to devise a rigorous gold standard that is used by all trial eligibility matching algorithms to evaluate and compare results and to create more generalizable methods. Furthermore, sharing the annotations of each study to an open-source can promote building larger training sets[16] among hospitals that store free-text patient information similarly or use the same electronic health record system. Some trial matching systems use machine learning and natural language processing, such as Deep 6 AI, Mendel.ai, Antidote, Smart Patients, and Synergy. However, there is a paucity in the studies that directly compare such tools, and abstracts and limited datasets limit analyses in this area.[9] As public data sets are more readily available, the initial hurdle of obtaining training data will be overcome and promote more active use of NLP[6] in clinical research.

Conclusion

Ground truth is needed in training, testing and evaluation steps of the development of NLP tools and is essential in achieving high robustness. However, in clinical NLP, there has been heterogeneity in the definition of ground truth which prevented an objective interpretation of the performance metrics reported and comparisons between different tools. The proposed methods will standardize the definition of ground truth used in clinical NLP in both small and large scale studies. This will be essential to creating generalizable and comparable clinical NLP tools.

Acknowledgement

The authors declare that there is no conflict of interest.

References

[1] Daugherty C, Ratain MJ, Grochowski E, et al. Perceptions of cancer patients and their physicians involved in phase I trials [published correction appears in J Clin Oncol 1995 Sep;13(9):2476]. *J Clin Oncol*. 1995;13(5):1062-1072. doi:10.1200/JCO.1995.13.5.1062

[2] Godskesen T, Hansson MG, Nygren P, Nordin K, Kihlbom U. Hope for a cure and altruism are the main motives behind participation in phase 3 clinical cancer trials. *Eur J Cancer Care (Engl)*. 2015;24(1):133-141. doi:10.1111/ecc.12184

[3] Moorcraft SY, Marriott C, Peckitt C, et al. Patients' willingness to participate in clinical trials and their views on aspects of cancer research: results of a prospective patient survey. *Trials*. 2016;17:17. Published 2016 Jan 9. doi:10.1186/s13063-015-1105-3

[4] Chen L, Gu Y, Ji X, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. *J Am Med Inform Assoc*. 2019;26(11):1218-1226. doi:10.1093/jamia/ocz109

[5] Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc*. 2019;26(4):294-305. doi:10.1093/jamia/ocy178

[6] Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform*. 2017;73:14-29. doi:10.1016/j.jbi.2017.07.012

[7] Dumitrache A, Aroyo L, Welty C. Crowdsourcing ground truth for medical relation extraction. 2017; *arXiv preprint arXiv:1701.02185*.

[8] Somashekhar SP, Sepúlveda MJ, Puglielli S, et al. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol*. 2018;29(2):418-423. doi:10.1093/annonc/mdx781

[9] Alexander M, Solomon B, Ball DL, et al. Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. *JAMIA Open*. 2020;3(2):209-215. Published 2020 May 1. doi:10.1093/jamiaopen/ooaa002

[10] Hassanzadeh H, Karimi S, Nguyen A. Matching patients to clinical trials using semantically enriched

document representation. *J Biomed Inform*. 2020;105:103406. doi:10.1016/j.jbi.2020.103406

[11] Beck JT, Rammage M, Jackson GP, et al. Artificial Intelligence Tool for Optimizing Eligibility Screening for Clinical Trials in a Large Community Cancer Center. *JCO Clin Cancer Inform*. 2020;4:50-59. doi:10.1200/CCI.19.00079

[12] Zhang K, Demner-Fushman D. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. *J Am Med Inform Assoc*. 2017;24(4):781-787. doi:10.1093/jamia/ocw176

[13] Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. *Int J Med Inform*. 2019;129:13-19. doi:10.1016/j.ijmedinf.2019.05.018

[14] Liang JJ, Tsou CH, Devarakonda MV. Ground Truth Creation for Complex Clinical NLP Tasks - an Iterative Vetting Approach and Lessons Learned. AMIA Jt Summits Transl Sci Proc. 2017;2017:203-212. Published 2017 Jul 26.

[15] Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc*. 2010;17(5):519-523. doi:10.1136/jamia.2010.004200

[16] Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*. 2011;18(5):540-543. doi:10.1136/amiajnl-2011-000465

Address for Correspondence

Jeongeun (Jane) Kim

Email: jkim@hms.harvard.edu