Article published online: 2022-06-02

Natural Language Processing: from Bedside to Everywhere

Eiji Aramaki[1], Shoko Wakamiya[1], Shuntaro Yada[1], Yuta Nakamura[2]
[1] Nara Institute of Science and Technology (NAIST), Nara, Japan
[2] Division of Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Summary

Objectives: Owing to the rapid progress of natural language processing (NLP), the role of NLP in the medical field has radically gained considerable attention from both NLP and medical informatics. Although numerous medical NLP papers are published annually, there is still a gap between basic NLP research and practical product development. This gap raises questions, such as what has medical NLP achieved in each medical field, and what is the burden for the practical use of NLP? This paper aims to clarify the above questions.

Methods: We explore the literature on potential NLP products/services applied to various medical/clinical/healthcare areas.

Results: This paper introduces clinical applications (bedside applications), in which we introduce the use of NLP for each clinical department, internal medicine, pre-surgery, post-surgery, oncology, radiology, pathology, psychiatry, rehabilitation, obstetrics, and gynecology. Also, we clarify technical problems to be addressed for encouraging bedside applications based on NLP.

Conclusions: These results contribute to discussions regarding potentially feasible NLP applications and highlight research gaps for future studies.

Keywords

Natural language processing, medical application, chatbot, randomized controlled trial, social media

Yearb Med Inform 2022:243-53
http://dx.doi.org/10.1055/s-0042-1742510

1 Introduction

Electronic health/medical records (referred to as EHR in this study) are rapidly replacing paper-based records in hospitals worldwide. Natural language processing (NLP) techniques have gained importance in the medical field. Because NLP is a hot topic in computer science, the number of medical NLP studies is increasing each year dramatically.

Despite the large number of studies, only a few practical studies have validated medical NLP applications in real-world settings. Studies using randomized controlled trials (RCTs), which have the highest medical evidence, are rare. In the PubMed search for "NLP" + "RCT" or "Clinical trial," we could find few studies only [1–4]. Instead of RCT studies, several studies employed a retrospective study using EHR big data: screening of diseases, case classification, incident detection, etc. [5–8]. However, unlike medical image software, these systems have not been commercialized as a product. A similar trend can be observed in the approved applications of the Food and Drug Administration (FDA) as artificial intelligence (AI) systems[1]. Most were audiology devices, and no medical systems related to NLP were found.

In summary, NLP has been actively studied, but there is still a gap between basic research and practical product development. This raises several questions, including what has medical NLP achieved in each medical field, and what is the burden for practical use of NLP? To clarify these questions, this study investigates what clinical/medical NLP has achieved in different clinical/medical fields.

This review aims to provide a guide for the NLP specialist who does not know medical informatics well enough. The scope of this paper is related to studies that have the potential to directly contribute to daily clinical practice, which we call bedside applications, consisting of internal medicine, pre-surgery, post-surgery, oncology, radiology, pathology, psychiatry, rehabilitation, obstetrics, and gynecology, etc. This paper introduces existing ready-to-use systems used in the above fields and summarizes its current methodology and performance. Finally, we mention future potential NLP applications not only for hospital use but also for patient use.

2 Bedside Applications

We provide an overview of how far NLP can be applied to outpatient and inpatient diagnosis, treatment, or management in each department. Historically, shared tasks have been one of the effective ways for researchers to drive fundamental innovations in the clinical NLP [9]. This is a competitive platform where organizers present a technically challenging and clinically meaningful task along with the dataset, gold standards, and evaluation criteria. In the early days, simple tasks were chosen, such as classifying patient records based on smoking status [10]. These days, shared tasks deal with far more complex problems, such as temporal relationship recognition among

[1] https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

clinical events in discharge summaries [11], risk factor identification in longitudinal series of progress notes [12], and clinical decision support [13–15]. Over time, reproducibility of solutions and techniques found in shared tasks have been demonstrated by researchers, which has promoted advancements in clinical NLP.

We surveyed how far NLP applications have been proven to be replicable in real-world clinical practice. We made no limitations on hospital departments in searching publications. We referred to (i) reviews and systematic reviews published in 2017 or later and (ii) original research articles published in 2020 or later on NLP applications for each hospital department. We searched PubMed for publications using the keyword "natural language processing" for reviews and systematic reviews, and "natural language processing", and a hospital department name together for original research articles. Because this article is not a systematic review, we focused on studies that can directly contribute to daily clinical practice. Although NLP is also helpful in research-oriented applications, such as cohort building with patient identification or phenotyping [16], evidence generation using clinical free-text [17–19], or semi-automation of meta-analysis [20] and systematic review [21–23], these are beyond the scope of this article.

2.1 Applications in Different Departments

NLP-based technology has enabled information extraction (IE) from various unstructured free-text documents such as clinic letters, progress notes, discharge summaries, and test reports. This technology can improve care quality in multiple departments, which has been demonstrated mainly in retrospective studies and sometimes in prospective studies [24–27]. NLP performance has also been validated in multicenter studies [28, 29]. See also Table 1 for details of the NLP systems introduced below.

Internal Medicine

NLP aids in the prevention, early diagnosis, treatment, and prognostic prediction of a wide range of diseases, such as cardiovascular, endocrine, metabolic, hepatobiliary, and neurological diseases [30].

(i) Disease prevention. NLP can identify risk factors, estimate risk, or predict events of disease development or readmissions [12, 31, 32]. Wang et al. automatically calculated CHA_2DS_2-VASc and HAS-BLED, the risk scores for the cerebral stroke of atrial fibrillation patients, by a rule-based approach. They also identified patients with a high risk of cerebral stroke with positive predictive values of 0.92–1.00 [33]. Buchan et al. analyzed clinical notes of patients without a history of coronary artery disease (CAD) with named entity recognition (NER) and support vector machine (SVM), and identified patients with later development of CAD with F1-score of 0.774 [34];

(ii) Early diagnosis. NLP can help clinicians recognize diseases out of their specialty that might otherwise be misdiagnosed or overlooked without proper transfer. Chase et al. achieved area under a receiver operating characteristic curve (ROC-AUC) of 0.94 in classifying patients with and without multiple sclerosis using NER and Naïve Bayes classifiers. They also identified patients suspected of undiagnosed multiple sclerosis [35];

(iii) Treatment support. Clinical decision support tools to summarize patient clinical information and suggest treatment are beginning to be realized. Seol et al. integrated a clinical decision support tool into the EHR system for pediatric asthma outpatients, which warns of the risk of acute exacerbation and recommends an optimal treatment plan based on free-text and structure data in the EHR [25]. An RCT demonstrated improvement of patient outcomes and significantly reduced physicians' workload for manual chart review.

Pre-surgery

NLP has the potential to aid in identifying clinical conditions of preoperative, perioperative, and postoperative patients [36, 37]. In preoperative settings, NLP can (i) evaluate surgical indications and (ii) reduce the workload of preoperative assessment. Wissel et al. implemented an automatic NLP scoring system in the EHR system that identifies epileptic outpatients with indications of surgery with SVM. The system achieved ROC-AUC of 0.79 in recommending operation [24]. Fonferko-Shadrach et al. developed an NLP system to review clinic letters and automatically extract symptoms, diagnosis, and medication history of preoperative patients. The system was based on an existing entity linking tool and demonstrated F1-score of 0.911 [38].

Post-surgery

Perioperatively and postoperatively, NLP contributes to continuous quality improvement efforts. NLP can identify complications and their details in unstructured free-text clinical records, even if they are not codified with ICD-10 (International Classification of Diseases -10th revision) [29, 39]. Bucher et al. identified surgical site infections (SSIs) with an NLP pipeline that parses and extracts information from clinical notes reaching ROC-AUC of 0.912. The system also determined SSI subgroups based on the depth, the wound condition, and the outcome [29]. Furthermore, surgical outcomes can also be automatically extracted from unstructured free-text using NLP, which aids labor-intensive manual chart review. In orthopedics, hip dislocation after total hip arthroplasty can be detected [40]. Tibbo et al. developed an NLP system to automatically determine Vancouver classification of periprosthetic femur fractures with the sensitivity of 0.786 and specificity of 0.948 [41].

Oncology

Oncology is another department where NLP plays an important role [30, 42].

(i) IE and cancer registration. NLP helps information retrieval on genetic, histological, and clinical characteristics of cancer, which is essential in clinical decision making and surveillance for effective public health interventions [43, 44]. The information includes histological type, differentiation, Ki-67 index, TNM (classification of malignant tumors) staging, test findings, treatment, family history, and performance status. Benjamin et al. automatically extracted quantitative information of biomarkers

from breast cancer pathology reports. They achieved an accuracy of 0.98 with a rule-based approach on top of an existing NER tool MetaMap [45, 46];

(ii) Clinical decision support. Precision medicine is a tailor-made clinical practice considering individual patient demographics and cancer genetic characteristics. NLP can recommend optimal treatment plans by searching biomedical articles and clinical trial repositories using patient information as a query [13–15, 47]. Li et al. released a chatbot-style open access clinical decision support tool [48].

Radiology

NLP can contribute to multiple stages of the radiological clinical workflow [49–51].

(i) Patient safety. NLP can help screen patients for contraindications to diagnostic imaging. Valtchinov et al. identified implants with contraindication to magnetic resonance imaging (MRI) in clinical notes with accuracies of 0.83–0.91 with NER [52];

(ii) Imaging protocol recommendation. NLP can determine the use of contrast agents or optimal imaging protocols based on free-text in ordering comments or clinical records [53–56]. Chillakuru et al. developed a machine learning-based NLP system to recommend the use of contrast agents for brain and spinal MRI with accuracies of 0.83–0.85, of which an online demo is available. The system is based on term frequency-inverse document frequency vectorization, Gradient Boosting Decision Tree (GBDT), word embeddings, and shallow neural networks [54]. Some other scan optimization tools are commercially available [55];

(iii) Automated radiology reporting. As the workload of diagnostic radiologists rapidly grows [57], automated radiology report generation in cooperation with computer vision AI is attracting attention [58]. Most studies have dealt with chest X-rays thus far, and further application to computed tomography (CT), MRI, and nuclear medicine is expected;

(iv) Surveillance. Radiology reports sometimes point out incidental findings. NLP can help prevent such findings from being missed by the attending physician by automatically sending alerts [49–51].

Pathology

NLP is helpful for both pathologists, whose responsibility is increasing in the era of personalized medicine, and clinicians, who refer to the diagnosis for treatment planning.

(i) Support diagnosis. NLP can support pathologists by providing a better computer-based image retrieval system incorporating pathology reports [59] or by automated pathology reporting [60];

(ii) Support clinical practice. Information on pathological diagnosis is used afterward by clinicians for better treatment strategy. NLP helps convert unstructured pathology reports into a structured form [45, 57, 61]. Kim et al. automatically extracted descriptions of a specimen, procedure, and pathologic diagnosis from pathology reports regardless of clinical departments. Their deep learning-based system, which uses Bidirectional Encoder Representations from Transformers (BERT), achieved accuracies of 0.9795–0.9839 [57, 62]. At a more fine-grained level, Odisho et al. extracted seventeen types of information from prostate cancer pathology reports and achieved a weighted F1-score of 0.972 for categorical data and a mean accuracy of 0.930 for numerical data. They applied document classification with convolutional neural network (CNN) to categorical data and token classification with random forest to numerical data [61].

Psychiatry

In psychiatry, NLP can be used for IE from unstructured EHR and speech analysis on patient speech data [63, 64]. NLP can help in the screening, early diagnosis, or severity estimation of various diseases such as depression [63], bipolar disorder [65], dementia [66–68], psychosis [69, 70], and schizophrenia [71]. Dai et al. showed that NLP automatically diagnosed psychiatric diseases with free-text discharge summaries. Their system achieved a micro F1-score of 0.584 using multiple classifiers based on pre-trained Robustly Optimized BERT pretraining Approach (RoBERTa) models [72, 73]. More fundamentally, NLP can contribute to psychiatric diagnostics. The Research Domain Criteria (RDoC), a potential counterpart of the Diagnostic and Statistical Manual of Mental Disorders (DSM), aims to integrate brain research knowledge into psychiatric disease classification [74], for which NLP shared tasks were held in 2016 and 2019 [75, 76].

Rehabilitation

NLP is used in speech therapy by incorporating it into electronic devices for augmentative and alternative communication (AAC) [77, 78]. Moreover, NLP has the potential to better unite the entire rehabilitation into the healthcare process by enabling the integration of the International Classification of Functioning, Disability, and Health (ICF) into EHRs, although there are still problems to overcome [79].

Obstetrics and Gynecology

Publications on bedside NLP applications were found in obstetrics and gynecology, although limited in number. Moon et al. showed the effectiveness of a rule-based NLP approach to highlight information discrepancies on surgical history due to misinterpretation during hospital transfer or improper copy and paste [80]. Sterckx et al. developed a birth risk prediction system to support preterm birth treatment, which was based on GBDT. NER-based features improved prediction performance when combined with structured data, with F1-score of birth prediction within 24 hours over 0.80 [81]. Barber et al. used NLP for prognostic prediction of ovarian cancer surgery, where postoperative readmission within 30 days was predicted with ROC-AUC of 0.70 using preoperative CT radiology reports [82].

Other Departments

NLP application is limited in ophthalmology and anesthesiology, where most AI systems are devoted to automated image diagnosis

[83] or intraoperative monitoring with numerical data [84]. However, some studies combine NLP for unstructured free-text documents and AI for structured EHR data to predict patient prognosis [85]. NLP also has the potential to automatically pick up patient risk factors preoperatively.

As indicated above, NLP can improve the quality and efficiency of bedside clinical practice mainly by IE from unstructured free-text for various departments and diseases, a part of which has already been put to practical use.

2.2 Cross-cutting Applications

Some NLP applications are not limited to specific hospital departments but can be helpful widely. We introduce such applications in this subsection.

Text Simplification

Clinical texts can sometimes be difficult for patients or clinicians in other departments due to jargon or abbreviations. Automated text simplification with NLP can improve both patient-staff and staff-staff communication [86, 87]. Moen et al. developed an NLP system to suggest replacements for abbreviations in Finnish clinical texts that are difficult for patients. The system achieved top-1 accuracy of 0.3464 with an unsupervised approach using cosine similarity of word embeddings [87].

Writing Support

Writing support with NLP can solve more fundamental problems that illegible clinical texts often result from a shortage of time of healthcare professionals for documentation.

(i) Auto-completion. Auto-completion is a real-time suggestion of the next word or clinical concept while a healthcare professional writes a clinical document. Gopinath et al. developed an auto-completion system for the emergency department that suggests clinical conditions, symptoms, medications, and laboratory test items during the documentation of progress notes. The system reduced the keystroke burden by 67% [88];

(ii) Auto-structuring. Some clinical documents such as progress notes or nursing notes are required to be in a structured form. NLP allows healthcare professionals to write such documents in an unstructured narrative by automatic editing and structuring. Moen et al. structured Finnish nursing notes into paragraphs whose headings were selected from standardized taxonomy with an accuracy of 0.71 using a Long Short-Term Memory (LSTM)-based sentence classification [89]. Furthermore, patient-staff conversations can be automatically structured once transcribed [90, 91];

(iii) Digital scribe. Digital scribe is different from dictation but similar to auto-structuring except for using voice input. That is, clinicians have only to record an outpatient conversation with some additional voice command, and the NLP system analyzes and summarizes the conversation and converts it into a clinical document in a predefined format [92–95]. Wang et al. developed a digital scribe system, which was 2.17–3.12 times faster than typing and dictation during patient encounter documentation [95].

3 Problems to be Addressed

3.1 Standard Annotation Schemes

Most NLP-based IE techniques adopted in the studies we referred to thus far use supervised machine learning, which requires high-quality, large datasets for training. Creating such datasets relies on manual annotation and thus increases the cost.

The formats and conventions of writing clinical documents differ not only in document types (e.g., EHRs, radiology reports, and nursing notes), but also in hospitals, departments, and even individual doctors. This textual diversity requires medical NLP researchers to create dedicated corpora for different applications by designing distinct annotation schemes. For instance, doctors often write disease name abbreviations in EHRs owing to the nature of personal note-taking, while radiology reports contain slightly more standardized terms because they are exchanged between diagnosing doctors and radiologists. Distributions of the appearing clinical terms in different types of clinical notes of different departments also deviate substantially, leading to uneven performance even when using an identical model architecture [96].

To adapt for a wide range of clinical note types with a single annotation scheme, some studies propose general-purpose annotation guidelines that define popular medical entities (e.g., diseases, drugs, tests, remedies, and body parts), as well as semantic relationships among them (e.g., "a medicine 'is-subscribed-for' a disease" and "a symptom 'was-found-in' an anatomical part") [96–99]. However, this approach increases the complexity of the resulting annotation schemes, making training annotators expensive. One guideline of such schemes has more than 30 pages [100]; a temporal IE corpus provides a 63 pages-long guideline document [101].

The complexity of annotation schemes can also generate ambiguous boundaries between multiple entity types. For example, a general-purpose corpus [99] defines 'Disease' entity and 'Signs or Symptoms' entity separately, the inter-annotator agreement of which was relatively low probably because of the annotators' confusion.

3.2 Task Formulation

There are always several ways to formulate a medical/clinical problem into an NLP task. The difference in task formulation affects overall performance and how to create an annotated corpus. Careful design of an NLP task setting translated from clinical needs matters. Taking adverse drug event (ADE) detection as an example, we have at least three options in its task formulation: NER, relation extraction (RE), and text classification. We represent these different approaches in Figure 1. The example sentence implies that a medication "nivolumab" prescribed for a "laryngeal cancer" adversely caused "liver damage." As we mentioned below, each approach has its own benefits and drawbacks. This trade-off suggests that we must carefully design NLP approaches against given medical/clinical IE issues.

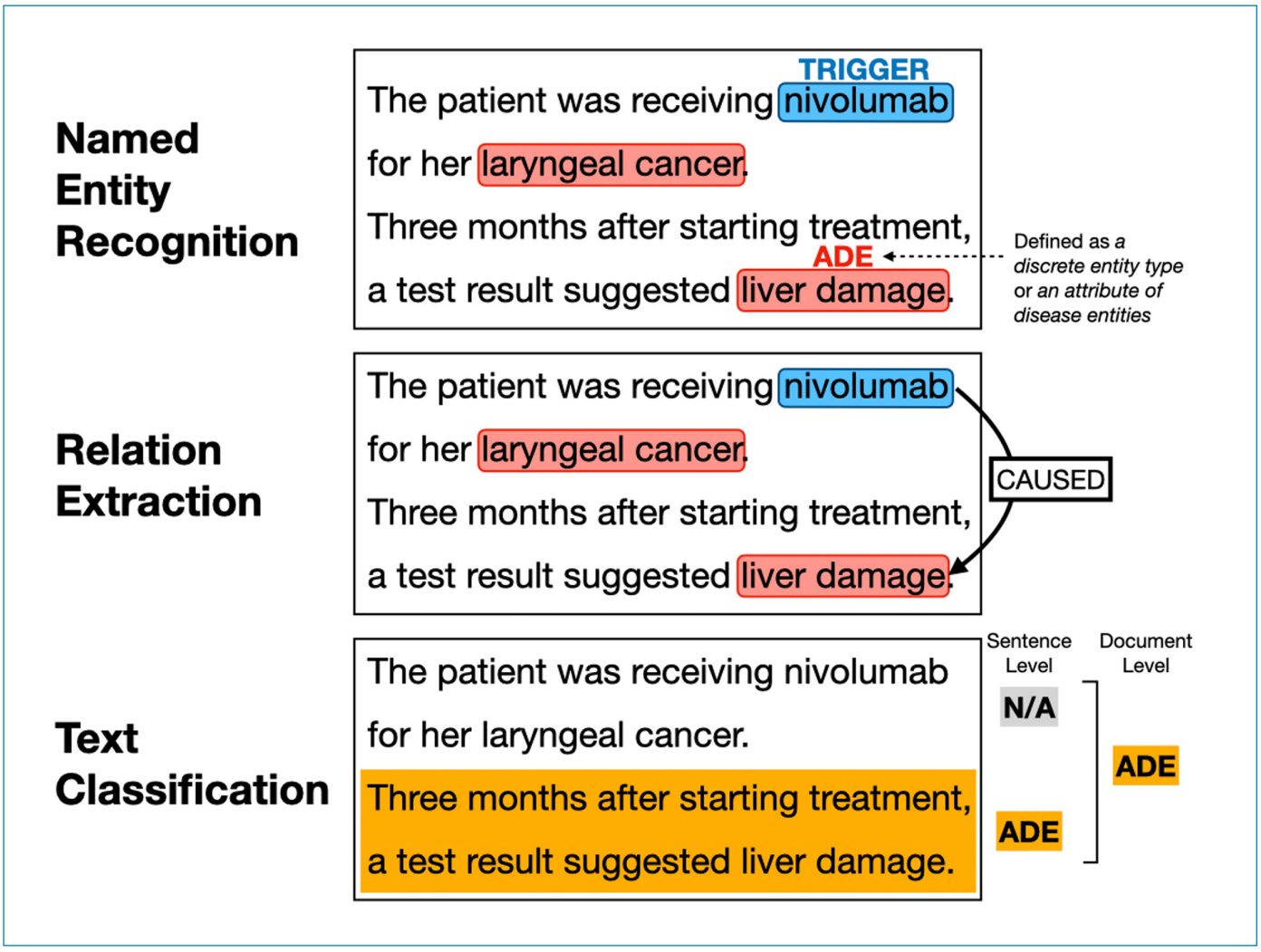

Fig. 1 Different task formulations for the same task (a case of the adverse drug event task).

Named Entity Recognition

One way of identifying an ADE is to label which disease entities were adversely caused by medication. We can adopt NER approaches, e.g., by directly labeling "ADE" entities [102–104]. In our example (the top row of Figure 1), this approach distinguishes "liver damage" as an ADE from a non-ADE disease entity "laryngeal cancer."

Another approach is to put a value "ADE" as an attribute to corresponding disease entities that are already labeled by the standard medical NER. In medical NER, some attributes can be assigned to an entity type, such as factuality (whether or not a disease was found in the patient) and schedule (when a medication was subscribed) [96, 105, 106]. In a shared-task workshop called Real-MedNLP[2], Subtask 3 ADE proposes such a task formulation, where the medication and disease/symptom entities found in a document are to be labeled ADE TRIGGER or ADE, respectively.

Although these simple approaches do not encode the information about which drug caused the adverse symptom (i.e., causal ADE relationships), it still works for initial screening. Probably because the models need to recognize longer context to detect ADE entities, they perform relatively lower (around 0.6 F1-score [104, 107]) than typical disease name recognition models (around 0.9 F1-score in the BC5CDR dataset [108]).

Relation Extraction

ADE detection tends to be defined as RE [103, 104, 107, 109] so that the causality information of possible ADEs is directly encoded. In our example (the middle row

[2] https://sociocom.naist.jp/real-mednlp/

of Figure 1), the drug entity "nivlumab" should be connected to "liver damage" by an ADE-causing relation ("CAUSED"). Additionally, the detail of medication treatments is often annotated, i.e., labeling drug-attribute relations from a drug entity to the expressions such as its amount or frequency of prescription.

However, it is not trivial even for professional clinicians to decide if a disorder written in a document was certainly caused by some drugs or not, which may result in difficulty in annotations [107]. In fact, the performance in detecting ADE relations, which distributes around 0.5 F1-score, were substantially low in comparison to drug-attribute relation extraction, most models of which achieved around 0.9 F1-score [103, 107, 109].

Text Classification

Another simple approach to ADE detection is classification-based IE, which detects ADE information mentioned in a document by sentence- [110, 111] or document-level [111] classification. For instance, Ujiie et al. [111] proposed a machine learning-based method to first classify each sentence of case reports into ADE-suggesting or not, and then to identify the documents that report any ADEs based on the sentence-level classification results. In our case (the bottom row of Figure 1), the second sentence is to be marked as ADE-suggesting since it mentions an ADE ("liver damage"), and hence the whole document containing the two sentences is to be labeled ADE-reporting.

This coarse-grained approach allows end-users who report ADEs from clinical documents to investigate the position in a document that suggests potential ADEs. The document-level classification seems to work better than sentence-level classification (around 0.5 vs 0.8 F1-score in [111]), probably due to the difficulty in inter-sentence relation understanding.

3.3 Real-time Nature, UI, UX of NLP

Despite its potential, the effectiveness of NLP applications has rarely been prospectively examined except for a few studies such as decision support for surgery candidacy [24]. There is a huge gap between retrospective studies and prospective studies. To break this out, a real-time NLP platform including a clinician-friendly graphical interface [25,112] is required.

4 Future NLP and Conclusions

Sections 2 and 3 described the clinical NLP systems in the hospital. Beyond its use in hospitals, NLP applications can be combined with a variety of smart devices, such as smartphones, smart speakers, and smartwatches. In the final part of this review, we pick up emerging out-of-hospital NLP applications that will grow potentially in the near future. Their core concepts of services are twofold: (1) for the patient and (2) for medical staff.

Peer support and conversation agents are core NLP targets for patients. Peer support is based on human-to-human communication. Nowadays, direct human communication has been gradually replaced by virtual communication. Rouzfarakh et al., for example, formed a WhatsApp peer support group for burn patients to share their experiences [113]. Zhang et al. developed a WeChat platform for parents of children with congenital heart diseases [114]. Yonek et al. performed a Facebook-based RCT for tobacco and heavy alcohol use [115]. Yang et al. explored the effect of WeChat follow-up management on improving parents' mental status and quality of life (QoL) in premature newborns with patent ductus arteriosus [116]. Thus, these previous studies focus more on forming virtual communication spaces where one can connect with peers and exploring their effectiveness without NLP techniques. As the next step, NLP would be applied for peer recommendation, communication facilitation support, effectiveness measurement of peer support, etc.

Instead of human conversational agents, NLP systems (conversation agents or chatbots) can provide mental encouragement to patients. Conversation agents have been developed for depression patients [117] and smokers [118], while some other agents are dedicated to promoting physical activity, a healthy diet [119], communication support for children with autism spectrum disorder [120], QoL control for inflammatory bowel disease (IBD) [121]. A clinical issue in such chatbot development is how to ensure patient safety [122]. To deal with this problem, new solutions are explored. For example, a system named Addiction-Comprehensive Health Enhancement Support System [123] implemented a panic button: if the patient pressed it, the system sends an emergency message to pre-registered contact people.

For the hospital, education and navigation are the main NLP targets. Communication skill training is a typical example for both doctors [124] and nurses [125]. Medical navigation, not only geographic but also information-oriented, is useful in medical applications. A successful example is to provide relevant information inside clinical departments [126]. Chu et al. developed a Question-and-Answer (QA) system for hospital staff to inform the location of mobile medical equipment (electrocardiography machines), moving around a hospital [127].

To conclude this paper, we refer to the first two questions in Section 1: what has medical NLP achieved in each medical field, and what is the burden for practical use of NLP? On the one hand, NLP-powered approaches have already been applied to most bedside needs. The performance of such approaches reached around 0.9 ROC-AUC, demonstrating the "in-vitro" feasibility of NLP for bedside applications. On the other hand, we observed several limitations in real-world use of NLP: too much variety of corpus-annotation schemes and task formulation lead to low portability of existing solutions; and lack of user-interface/experience evaluations concerns clinicians about "in-vivo" usability. The potential coverage of medical NLP is yet broader than direct bedside applications, as introduced in this section. Realization of successful medical NLP applications may need a much larger-scale, interdisciplinary collaboration involving bedside staff, patients, UI/UX scholars, wearable Internet of Things devices, and NLP researchers.

References

1. Agurto C, Cecchi GA, Norel R, Ostrand R, Kirkpatrick M, Baggott MJ, et al. Detection of acute 3,4-methylenedioxymethamphetamine (MDMA) effects across protocols using automated natural language processing. Neuropsychopharmacology.2020 Apr;45(5):823-2.
2. Pestian JP, Grupp-Phelan J, Bretonnel Cohen K, Meyers G, Richey LA, Matykiewicz P, et al. A Controlled Trial Using Natural Language Processing to Examine the Language of Suicidal Adolescents in the Emergency Department. Suicide Life Threat Behav 2016 Apr;46(2):154-9.
3. Boyé M, Grabar N, Thi Tran M. Contrastive conversational analysis of language production by Alzheimer's and control people. Stud Health Technol Inform 2014;205:682-6.
4. Xia Z, Secor E, Chibnik LB, Bove RM, Cheng S, Chitnis T, et al. Modeling disease severity in multiple sclerosis using electronic health records. PLoS One 2013 Nov 11;8(11):e78927.
5. Mendonça EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform 2005 Aug;38(4):314-21.
6. Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm Bowel Dis 2013 Jun;19(7):1411-20.
7 . Shiner B, Neily J, Mills PD, Watts BV. Identification of Inpatient Falls Using Automated Review of Text-Based Medical Records. J Patient Saf 2020 Sep;16(3):e174-e178.
8. Salmasian H, Freedberg DE, Abrams JA, Friedman C. An automated tool for detecting medication overuse based on the electronic health records. Pharmacoepidemiol Drug Saf 2013 Feb;22(2):183-9.
9. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner Ö. 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 Sep-Oct;18(5):540-3.
10. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008 Jan-Feb;15(1):14-24.
11. Sun W, Rumshisky A, Uzuner Ö. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013 Sep-Oct;20(5):806-13.
12. Stubbs A, Kotfila C, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform 2015 Dec;58 Suppl(-Suppl):S67-S77.
13. Roberts K, Demner-Fushman D, Voorhees EM, Bedrick S, Hersh WR. Overview of the TREC 2020 Precision Medicine Track. Text Retr Conf 2020 Nov;1266.
14. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, et al. Overview of the TREC 2019 Precision Medicine Track. Text Retr Conf 2019 Nov;1250.
15. Robert K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, et al. Overview of the TREC 2017 Precision Medicine Track. Text Retr Conf 2017 Nov;26.
16. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural Language Processing for EHR-Based Computational Phenotyping. IEEE/ACM Trans Comput Biol Bioinform 2019 Jan-Feb;16(1):139-53.
17. Ross EG, Shah N, Leeper N. Statin Intensity or Achieved LDL? Practice-based Evidence for the Evaluation of New Cholesterol Treatment Guidelines. PLoS One 2016 May 26;11(5):e0154952.
18. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLoS One 2013 May 23;8(5):e63499.
19. Smith DH, Johnson ES, Russell A, Hazlehurst B, Muraki C, Nichols GA, et al. Lower visual acuity predicts worse utility values among patients with type 2 diabetes. Qual Life Res 2008 Dec;17(10):1277-84.
20. Norman CR, Leeflang MMG, Porcher R, Névéol A. Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. Syst Rev 2019 Oct 28;8(1):243.
21. Gartlehner G, Wagner G, Lux L, Affengruber L, Dobrescu A, Kaminski-Hartenthaler A, Viswanathan M. Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. Syst Rev 2019 Nov 15;8(1):277.
22. Schmidt L, Olorisade BK, McGuinness LA, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: A living review protocol. F1000Res 2020 Mar 25;9:210.
23. Zimmerman J, Soler RE, Lavinder J, Murphy S, Atkins C, Hulbert L, et al. Iterative guided machine learning-assisted systematic literature reviews: a diabetes case study. Syst Rev 2021 Apr 2;10(1):97.
24. Wissel BD, Greiner HM, Glauser TA, Holland-Bouley KD, Mangano FT, Santel D, et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. Epilepsia 2020 Jan;61(1):39-48.
25. Seol HY, Shrestha P, Muth JF, Wi CI, Sohn S, Ryu E, et al. Artificial intelligence-assisted clinical decision support for childhood asthma management: A randomized clinical trial. PLoS One 2021 Aug 2;16(8):e0255261.
26. Zheng L, Wang Y, Hao S, Shin AY, Jin B, Ngo AD, et al. Web-based Real-Time Case Finding for the Population Health Management of Patients With Diabetes Mellitus: A Prospective Validation of the Natural Language Processing-Based Algorithm With Statewide Electronic Medical Records. JMIR Med Inform 2016 Nov 11;4(4):e37.
27. Wang Y, Luo J, Hao S, Xu H, Shin AY, Jin B, et al. NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. Int J Med Inform 2015 Dec;84(12):1039-47.
28. Tian Z, Sun S, Eguale T, Rochefort CM. Automated Extraction of VTE Events From Narrative Radiology Reports in Electronic Health Records: A Validation Study. Med Care 2017 Oct;55(10):e73-e80.
29. Bucher BT, Shi J, Ferraro JP, Skarda DE, Samore MH, Hurdle JF, et al. Portable Automated Surveillance of Surgical Site Infections Using Natural Language Processing: Development and Validation. Ann Surg 2020 Oct;272(4):629-36.
30. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform 2019 Apr 27;7(2):e12239.
31. Shi X, Hu Y, Zhang Y, Li W, Hao Y, Alelaiwi A, et al. Multiple disease risk assessment with uniform model based on medical clinical notes. IEEE Access 2016;4:7074–83.
32. Watson AJ, O'Rourke J, Jethwani K, Cami A, Stern TA, Kvedar JC, et al. Linking electronic health record-extracted psychosocial data in real-time to risk of readmission for heart failure. Psychosomatics. 2011 Jul-Aug;52(4):319-27.
33. Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc 2017 Mar 1;24(2):339-44.
34. Buchan K, Filannino M, Uzuner Ö. Automatic prediction of coronary artery disease from clinical narratives. J Biomed Inform 2017 Aug;72:23-32.
35. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak 2017 Feb 28;17(1):24.
36. Mellia JA, Basta MN, Toyoda Y, Othman S, Elfanagely O, Morris MP, et al. Natural Language Processing in Surgery: A Systematic Review and Meta-analysis. Ann Surg 2021 May 1;273(5):900-8.
37. Wyatt JM, Booth GJ, Goldman AH. Natural Language Processing and Its Use in Orthopaedic Research. Curr Rev Musculoskelet Med 2021 Dec;14(6):392-6.
38. Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford DV, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open 2019 Apr 1;9(4):e023232.
39. Selby LV, Narain WR, Russo A, Strong VE, Stetson P. Autonomous detection, grading, and reporting of postoperative complications using natural language processing. Surgery 2018 Dec;164(6):1300-5.
40. Borjali A, Magnéli M, Shin D, Malchau H, Muratoglu OK, Varadarajan KM. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation. Comput Biol Med 2021 Feb;129:104140.
41. Tibbo ME, Wyles CC, Fu S, Sohn S, Lewallen DG, Berry DJ, et al. Use of Natural Language Processing Tools to Identify and Classify Periprosthetic Femur Fractures. J Arthroplasty 2019 Oct;34(10):2216-9.
42. Sollini M, Bartoli F, Marciano A, Zanca R, Slart RHJA, Erba PA. Artificial intelligence and hybrid imaging: the best match for personalized med-

icine in oncology. Eur J Hybrid Imaging 2020 Dec 9;4(1):24.
43. Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. J Biomed Inform 2019 Dec;100:103301.
44. Tucker TC, Durbin EB, McDowell JK, Huang B. Unlocking the potential of population-based cancer registries. Cancer 2019 Nov 1;125(21):3729-37.
45. Holmes B, Chitale D, Loving J, Tran M, Subramanian V, Berry A, et al. Customizable Natural Language Processing Biomarker Extraction Tool. JCO Clin Cancer Inform 2021 Aug;5:833-41.
46. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21.
47. Xu J, Yang P, Xue S, Sharma B, Sanchez-Martin M, Wang F, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum Genet 2019 Feb;138(2):109-24.
48. Li J, Chen H, Wang Y, Chen MM, Liang H. Next-Generation Analytics for Omics Data. Cancer Cell 2021 Jan 11;39(1):3-6.
49. Pons E, Braun LM, Hunink MG, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology 2016 May;279(2):329-43.
50. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021 Jun 3;21(1):179.
51. Davidson EM, Poon MTC, Casey A, Grivas A, Duma D, Dong H, et al. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. BMC Med Imaging 2021 Oct 2;21(1):142.
52. Valtchinov VI, Lacson R, Wang A, Khorasani R. Comparing Artificial Intelligence Approaches to Retrieve Clinical Reports Documenting Implantable Devices Posing MRI Safety Risks. J Am Coll Radiol 2020 Feb;17(2):272-9.
53. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic Determination of the Need for Intravenous Contrast in Musculoskeletal MRI Examinations Using IBM Watson's Natural Language Processing Algorithm. J Digit Imaging 2018 Apr;31(2):245-51.
54. Chillakuru YR, Munjal S, Laguna B, Chen TL, Chaudhari GR, Vu T, et al. Development and web deployment of an automated neuroradiology MRI protocoling tool with natural language processing. BMC Med Inform Decis Mak 2021 Jul 12;21(1):213.
55. DSS Inc. Radiology Decision Support (RadWise®). Available from: https://www.dssinc.com/products/integrated-clinical-products/radwise-radiology-decision-support/
56. Letourneau-Guillon L, Camirand D, Guilbert F, Forghani R. Artificial Intelligence Applications for Workflow, Process Optimization and Predictive Analytics. Neuroimaging Clin N Am 2020 Nov;30(4):e1-e15.
57. Kim Y, Lee JH, Choi S, Lee JM, Kim JH, Seok J, et al. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Sci Rep 2020 Nov 20;10(1):20265.
58. Monshi MMA, Poon J, Chung V. Deep learning in generating radiology reports: A survey. Artif Intell Med 2020 Jun;106:101878.
59. Tizhoosh HR, Diamandis P, Campbell CJV, Safarpoor A, Kalra S, Maleki D, et al. Searching Images for Consensus: Can AI Remove Observer Variability in Pathology? Am J Pathol 2021 Oct;191(10):1702-8.
60. Pavlopoulos J, Kougia V, Androutsopoulos I. A survey on biomedical image captioning. In: Proceedings of the second workshop on shortcomings in vision and language; 2019. p.26-36.
61. Odisho AY, Park B, Altieri N, DeNero J, Cooperberg MR, Carroll PR, Yu B. Natural language processing systems for pathology parsing in limited data environments with uncertainty estimation. JAMIA Open 2020 Oct 14;3(3):431-8.
62. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019. p. 4171–86.
63. DeSouza DD, Robin J, Gumus M, Yeung A. Natural Language Processing as an Emerging Tool to Detect Late-Life Depression. Front Psychiatry 2021 Sep 6;12:719125.
64. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. J Med Internet Res 2021 May 4;23(5):e15708.
65. Jan Z, Ai-Ansari N, Mousa O, Abd-Alrazaq A, Ahmed A, Alam T, et al. The Role of Machine Learning in Diagnosing Bipolar Disorder: Scoping Review. J Med Internet Res 2021 Nov 19;23(11):e29749.
66. Graham SA, Lee EE, Jeste DV, Van Patten R, Twamley EW, Nebeker C, et al. Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. Psychiatry Res 2020 Feb;284:112732.
67. Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J Am Med Inform Assoc 2020 Nov 1;27(11):1784-97.
68. de la Fuente Garcia S, Ritchie CW, Luz S. Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. J Alzheimers Dis 2020;78(4):1547-74.
69. Corcoran CM, Mittal VA, Bearden CE, E Gur R, Hitczenko K, Bilgrami Z, et al. Language as a biomarker for psychosis: A natural language processing approach. Schizophr Res 2020 Dec;226:158-66.
70. Corcoran CM, Cecchi GA. Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. Biol Psychiatry Cogn Neurosci Neuroimaging 2020 Aug;5(8):770-9.
71. Ratana R, Sharifzadeh H, Krishnan J, Pang S. A Comprehensive Review of Computational Methods for Automatic Prediction of Schizophrenia With Insight Into Indigenous Populations. Front Psychiatry 2019 Sep 12;10:659.
72. Dai HJ, Su CH, Lee YQ, Zhang YC, Wang CK, Kuo CJ, et al. Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients. Front Psychiatry 2021 Jan 15;11:533949.
73. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint 2019; arXiv:1907.11692.
74. Cuthbert BN. The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. World Psychiatry 2014 Feb;13(1):28-35.
75. Uzuner Ö, Stubbs A, Filannino M. A natural language processing challenge for clinical records: Research Domains Criteria (RDoC) for psychiatry. J Biomed Inform 2017 Nov;75S:S1-S3.
76. Anani M, Kazi N, Kuntz M, Kahanda I. RDoC task at BioNLP-OST 2019. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks; 2019. p. 216-26.
77. Elsahar Y, Hu S, Bouazza-Marouf K, Kerr D, Mansor A. Augmentative and Alternative Communication (AAC) Advances: A Review of Configurations for Individuals with a Speech Disability. Sensors (Basel) 2019 Apr 22;19(8):1911.
78. Higginbotham DJ, Lesher GW, Moulton BJ, Roark B. The application of natural language processing to augmentative and alternative communication. Assist Technol 2011 Spring;24(1):14-24.
79. Maritz R, Aronsky D, Prodinger B. The International Classification of Functioning, Disability and Health (ICF) in Electronic Health Records. A Systematic Literature Review. Appl Clin Inform 2017 Dec 20;8(3):964-80.
80. Moon S, Carlson LA, Moser ED, Agnikula Kshatriya BS, Smith CY, Rocca WA, et al. Identifying Information Gaps in Electronic Health Records by Using Natural Language Processing: Gynecologic Surgery History Identification. J Med Internet Res 2022 Jan 28;24(1):e29015.
81. Sterckx L, Vandewiele G, Dehaene I, Janssens O, Ongenae F, De Backere F, et al. Clinical information extraction for preterm birth risk prediction. J Biomed Inform 2020 Oct;110:103544.
82. Barber EL, Garg R, Persenaire C, Simon M. Natural language processing with machine learning to predict outcomes after ovarian cancer surgery. Gynecol Oncol 2021 Jan;160(1):182-6.
83. Lin WC, Chen JS, Chiang MF, Hribar MR. Applications of Artificial Intelligence to Electronic Health Record Data in Ophthalmology. Transl Vis Sci Technol 2020 Feb 27;9(2):13.
84. Connor CW. Artificial Intelligence and Machine Learning in Anesthesiology. Anesthesiology 2019 Dec;131(6):1346-59.
85. Gaskin GL, Pershing S, Cole TS, Shah NH. Predictive modeling of risk factors and complications of cataract surgery. Eur J Ophthalmol 2016 Jun 10;26(4):328-37.
86. Mowery DL, South BR, Christensen L, Leng J, Peltonen LM, Salanterä S, et al. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2. J Biomed Semantics 2016 Jul 1;7:43.
87. Moen H, Peltonen LM, Koivumäki M, Suhonen H, Salakoski T, Ginter F, et al. Improving Layman

Readability of Clinical Narratives with Unsupervised Synonym Replacement. Stud Health Technol Inform 2018;247:725-9.
88. Gopinath D, Agrawal M, Murray L, Horng S, Karger D, Sontag D. Fast, Structured clinical documentation via contextual autocomplete. In: Proceedings of the Machine Learning for Healthcare Conference; 2020. p. 842–70.
89. Moen H, Hakala K, Peltonen LM, Matinolli HM, Suhonen H, Terho K, et al. Assisting nurses in care documentation: from automated sentence classification to coherent document structures with subject headings. J Biomed Semantics 2020 Sep 1;11(1):10.
90. Krishna K, Khosla S, Bigham JP, Lipton ZC. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics; 2021. p. 4958–72.
91. Zhang L, Negrinho R, Ghosh A, Jagannathan V, Hassanzadeh HR, Schaaf T, et al. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2021; 2021. p. 3693–712.
92. Blackley SV, Huynh J, Wang L, Korach Z, Zhou L. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. J Am Med Inform Assoc 2019 Apr 1;26(4):324-38.
93. Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. NPJ Digit Med 2018 Oct 16;1:58.
94. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med 2021 Mar 26;4(1):57.
95. Wang J, Lavender M, Hoque E, Brophy P, Kautz H. A patient-centered digital scribe for automatic medical documentation. JAMIA Open 2021 Feb 17;4(1):ooab003.
96. Yada S, Joh A, Tanaka R, Cheng F, Aramaki E, Kurohashi S. Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: starting from critical lung diseases. In: Proceedings of the 12th Language Resources and Evaluation Conference; 2020. p. 4565-72.
97. Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. J Biomed Inform 2009 Oct;42(5):950-66.
98. Patel P, Davey D, Panchal V, Pathak P. Annotation of a large clinical entity corpus. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018. p. 2033-42.
99. Campillos L, Deléger L, Grouin C, Hamon T, Ligozat AL, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). Language Resources and Evaluation 2018;52(2):571-601.
100. Yada S, Aramaki E, Tanaka R, Cheng F, Kurohashi S. Medical/Clinical text annotation guidelines. figshare. Book; 2021. Available from: https://doi.org/10.6084/m9.figshare.16418811.v2
101. Styler WF, Bethard S, Finan S, Palmer M, Pradhan S, De Groen PC, et al. Temporal annotation in the clinical domain. Transactions of the association for computational linguistics 2014;2:143-54.
102. Karimi S, Metke-Jimenez A, Kemp M, Wang C. Cadec: A corpus of adverse drug event annotations. J Biomed Inform 2015 Jun;55:73-81.
103. Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. In: Proceedings of the Tenth Text Analysis Conference 2017.
104. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner Ö. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 1;27(1):3-12.
105. Aramaki E, Yano K, Wakamiya S. MedEx/J: A One-Scan Simple and Fast NLP Tool for Japanese Clinical Texts. Stud Health Technol Inform 2017;245:285-8.
106. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc 2007 May-Jun;14(3):304-11.
107. Jagannatha A, Liu F, Liu W, Yu H. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). Drug Saf 2019 Jan;42(1):99-111.
108. Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford) 2016 Mar 19;2016:baw032.
109. Bayer S, Clark C, Dang O, Aberdeen J, Brajovic S, Swank K, et al. ADE Eval: An Evaluation of Text Processing Systems for Adverse Event Extraction from Drug Labels for Pharmacovigilance. Drug Saf 2021 Jan;44(1):83-94.
110. Negi K, Pavuri A, Patel L, Jain C. A novel method for drug-adverse event extraction using machine learning. Informatics in Medicine Unlocked 2019;17:100190.
111. Ujiie S, Yada S, Wakamiya S, Aramaki E. Identification of Adverse Drug Event-Related Japanese Articles: Natural Language Processing Analysis. JMIR Med Inform 2020 Nov 27;8(11):e22661.
112. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med 2019 Dec 17;2:130.
113. Rouzfarakh M, Deldar K, Froutan R, Ahmadabadi A, Mazlom SR. The effect of rehabilitation education through social media on the quality of life in burn patients: a randomized, controlled, clinical trial. BMC Med Inform Decis Mak 2021 Feb 22;21(1):70.
114. Zhang QL, Xu N, Huang ST, Chen Q, Cao H. WeChat-Assisted Preoperative Health Education Reduces Burden of Care on Parents of Children with Simple Congenital Heart Disease: a Prospective Randomized Controlled Study. Braz J Cardiovasc Surg 2021 Oct 17;36(5):663-9.
115. Yonek JC, Meacham MC, Ramo D, Delucchi K, Tolou-Shams M, Satre DD. The Relationship of E-Cigarette Use to Tobacco Use Outcomes Among Young Adults Who Smoke and Use Alcohol. J Addict Med 2021 Sep-Oct 01;15(5):421-4.
116. Yang B, Liu JF, Xie WP, Cao H, Chen Q. The effects of WeChat follow-up management to improve the parents' mental status and the quality of life of premature newborns with patent ductus arteriosus. J Cardiothorac Surg 2021 Aug 21;16(1):235.
117. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: A scoping review. Int J Med Inform 2019 Dec;132:103978.
118. Almusharraf F, Rose J, Selby P. Engaging Unmotivated Smokers to Move Toward Quitting: Design of Motivational Interviewing-Based Chatbot Through Iterative Interactions. J Med Internet Res 2020 Nov 3;22(11):e20251.
119. Zhang Y, Sun S, Galley M, Chen Y-C, Brockett C, Gao X, et al. DialoGPT: Large-scale generative pre-training for conversational response generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations. Association for Computational Linguistics; 2020. p. 270–8.
120. Cooper A, Ireland D. Designing a Chat-Bot for Non-Verbal Children on the Autism Spectrum. Stud Health Technol Inform 2018;252:63-8.
121. Zand A, Sharma A, Stokes Z, Reynolds C, Montilla A, Sauk J, Hommes D. An Exploration Into the Use of a Chatbot for Patients With Inflammatory Bowel Diseases: Retrospective Cohort Study. J Med Internet Res 2020 May 26;22(5):e15589.
122. Glowacki EM, Bernhardt JM, McGlone MS. Tailored texts: An application of regulatory fit to text messages designed to reduce high-risk drinking. Health Informatics J 2020 Sep;26(3):1742763.
123. Glass JE, McKay JR, Gustafson DH, Kornfield R, Rathouz PJ, McTavish FM, et al. Treatment seeking as a mechanism of change in a randomized controlled trial of a mobile health intervention to support recovery from alcohol use disorders. J Subst Abuse Treat 2017 Jun;77:57-66.
124. Reiswich A, Haag M. Evaluation of Chatbot Prototypes for Taking the Virtual Patient's History. Stud Health Technol Inform 2019;260:73-80.
125. Shorey S, Ang E, Yap J, Ng ED, Lau ST, Chui CK. A Virtual Counseling Application Using Artificial Intelligence for Communication Skills Training in Nursing Education: Development Study. J Med Internet Res 2019 Oct 29;21(10):e14658. Erratum in: J Med Internet Res 2019 Nov 26;21(11):e17064.
126. Lee H, Kang J, Yeo J. Medical Specialty Recommendations by an Artificial Intelligence Chatbot on a Smartphone: Development and Deployment. J Med Internet Res 2021 May 6;23(5):e27460.
127. Chu ET, Huang ZZ. DBOS: A Dialog-Based Object Query System for Hospital Nurses. Sensors (Basel) 2020 Nov 19;20(22):6639.

Correspondence to:
Eiji Aramaki
Nara Institute of Science and Technology (NAIST)
Nara, Japan
E-mail: aramaki@is.naist.jp

Table 1 Summary of bedside NLP application studies. BART = Bidirectional Auto-Regressive Transformer, BERT = Bidirectional Encoder Representations of Transformers, CNN = Convolutional Neural Network, EL = Entity Linking, GBDT = Gradient Boosting Decision Tree, LSTM = Long Short-Term Memory, NER = Named Entity Recognition, NN = Neural Network, RCT = Randomized Controlled Trial, RoBERTa = Robustly Optimized BERT Pretraining Approach, SVM = Support Vector Machine, T5 = Text-to-Text Transfer Transformer

Reference	Objectives	Study design	Target language	Corpus	NLP task	Method	Performance [95% confidence interval]
Wissel et al., 2020 [24]*	Detection of surgical candidates in epilepsy patients	Prospective study	English	Progress notes	Document classification	SVM	ROC-AUC 0.79 [0.62–0.96]
Seol et al., 2021 [25]*	Prevention of acute exacerbation of pediatric asthma with decision support tool	RCT	English	Progress notes	NER	Not applicable	Odds ratio 0.82 [0.374-1.96]
Zheng et al., 2016 [26]	Detection of patients with diabetes prior to structured coding	Prospective study	English	Progress notes, discharge summaries, and other clinical notes	Document classification	NER + Normalization + Random forest	ROC-AUC 0.929
Wang et al., 2015 [27]	Detection of chronic heart failure patients prior to structured coding	Prospective study	English	Progress notes, discharge summaries, and other clinical notes	Document classification	NER + Normalization + Random forest	ROC-AUC 0.919
Tian et al., 2017 [28]	Detection of deep vein thrombosis and pulmonary embolism	Retrospective study	English	Radiology reports	Document classification	Rule-based	Sensitivity 0.94 [0.88–0.97], Specificity 0.96 [0.95–0.97]
Bucher et al., 2020 [29]*	Surveillance of surgical site infections	Retrospective study	English	Operative reports, progress notes, nursing notes, radiology reports, and discharge summaries	Binary classification	Existing tool	ROC-AUC 0.912
Shi et al., 2016 [31]	Detection of cerebral infarction patients, pneumonia patients, and coronary artery disease patients	Retrospective study	Chinese	Progress notes and discharge summaries	Document classification	CNN	F1-score 0.934–0.966
Waston et al., 2011 [32]	Identification of psychosocial re-admission risk factor of heart failure patients	Retrospective study	English	EHR notes	Multivariate analysis	Rule-based feature extraction + Logistic regression	(Not applicable)
Wang et al., 2017 [33]	Detection of atrial fibrillation patients with high risk of cerebral infarction	Retrospective study	English	Clinical notes	Information extraction	Rule-based	Positive Predictive Value 0.92–1.00
Buchan et al., 2017 [34]	Prediction of developing coronary artery disease	Retrospective study	English	Clinical notes	Document classification	NER-based feature extraction + SVM	F1-score 0.774
Chase et al., 2017 [35]	Detection of multiple sclerosis patients	Retrospective study	English	EHR notes	Document classification	NER + EL + Naïve Bayes classification	ROC-AUC 0.94 [0.93–0.95]
Fonferko-Shadrach et al., 2019 [38]	Information extraction from clinical letters for epilepsy patients	Retrospective study	English	Clinic letters	Information extraction	Existing tool	F1-score 0.911
Selby et al., 2018 [39]	Surveillance of postoperative deep vein thrombosis and pulmonary embolism	Retrospective study	English	Radiology reports	Document classification	(Not applicable)	Sensitivity 0.851–0.900, Specificity 0.946–0.987
Borjali et al., 2021 [40]*	Surveillance of hip dislocation after total hip replacement	Retrospective study	English	Radiology reports and follow-up telephone notes	Document classification	LSTM, CNN	Kappa coefficient 0.97–1.00
Tibbo et al., 2019 [41]	Surveillance of periprosthetic femur fractures	Retrospective study	English	Operative reports and progress notes	Document classification	NER + EL	Sensitivity 1.000, Specificity 0.998
Holmes et al., 2021 [45]*	Information extraction of breast cancer biomarkers	Retrospective study	English	Pathology reports	Information extraction	NER + EL + Rule-based approach	Accuracy 0.98

* Papers published in 2020 or later.

Table 1 (continued) Summary of bedside NLP application studies. BART = Bidirectional Auto-Regressive Transformer, BERT = Bidirectional Encoder Representations of Transformers, CNN = Convolutional Neural Network, EL = Entity Linking, GBDT = Gradient Boosting Decision Tree, LSTM = Long Short-Term Memory, NER = Named Entity Recognition, NN = Neural Network, RCT = Randomized Controlled Trial, RoBERTa = Robustly Optimized BERT Pretraining Approach, SVM = Support Vector Machine, T5 = Text-to-Text Transfer Transformer

Reference	Objectives	Study design	Target language	Corpus	NLP task	Method	Performance [95% confidence interval]
Valtchinov et al., 2020 [52]*	Identification of implantable device posing MRI safety risks	Retrospective study	English	Radiology reports and other clinical notes	NER	Existing tool	Accuracy 0.83–0.91
Trivedi et al., 2018 [53]	Determination to use or not to use contrast enhancement for MRI imaging	Retrospective study	English	Free-text MRI protocols and indications	Document classification	Existing tool	Accuracy 0.832
Chillakuru et al., 2021 [54]*	Determination to use or not to use contrast enhancement for MRI imaging	Retrospective study	English	Free-text MRI protocols and indications	Document classification	TF-IDF vectorization + GBDT, Word embedding + shallow NN	Accuracy 0.8338–0.8543
Kim et al., 2020 [57]*	Information extraction from pathology reports	Retrospective study	English	Pathology reports	NER	BERT	Accuracy 0.9795–0.9839
Odisho et al., 2020 [61]*	Information extraction of histological characteristics of prostate cancer	Retrospective study	English	Pathology reports	Document classification and token classification	CNN, Random forest	Weighted F1-score 0.972 (document), Accuracy 0.930
Dai et al., 2020 [72]*	Automated diagnosis of multiple psychiatric diseases	Retrospective study	English and Chinese	Discharge summaries	Multilabel document classification	RoBERTa	Micro F1-score 0.584
Moon et al., 2022 [80]*	Recognition of gynecological surgical history	Retrospective study	English	Clinical notes	Multiclass classification	NER + Rule-based approach	Weighted F1-score 0.76
Sterckx et al., 2020 [81]*	Birth risk estimation	Retrospective study	English and Dutch	Clinical notes and structured data	Binary classification	GBDT	F1-score > 0.80
Barber et al., 2021 [82]*	Prediction of 30-day readmission after ovarian cancer surgery	Retrospective study	English	Radiology reports	Binary classification	(Not applicable)	ROC-AUC 0.70 [0.68–0.73]
Gaskin et al., 2016 [85]	Surveillance of postoperative complications of cataract surgery	Retrospective study	English	EHR notes and structured data	Binary classification	Random forest	ROC-AUC 0.62–0.84
Moen et al., 2018 [87]	Synonym replacement for better readability	Retrospective study	Finnish	Progress notes and nursing notes	Abbreviation Resolution	Word embedding + cosine similarity	Top-1 accuracy 0.3464
Gopinath et al., 2020 [88]*	Auto-completion of progress notes	Retrospective study	English	Progress notes and structured data	Recommen-dation	Rule-based entity type detection + Shallow NN or Bayes statistics	Keystroke reduction 67%
Moen et al., 2020 [89]*	Auto-structurization of nursing notes	Retrospective study	Finnish	Nursing notes	Sentence classification	LSTM	Accuracy 0.71
Krishna et al., 2021 [90]*	Automatic summarization of doctor-patient conversation	Retrospective study	English	Outpatient transcription	Summarization	BERT-LSTM classifier + T5	ROUGE-L 0.3838
Zhang et al., 2021 [91]*	Automatic summarization of doctor-patient conversation	Retrospective study	English	Outpatient transcription	Summarization	BART	ROUGE-L 0.3412

* Papers published in 2020 or later.