Natural Language Processing: from Bedside to Everywhere

Eiji Aramaki¹, Shoko Wakamiya¹, Shuntaro Yada¹, Yuta Nakamura²
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

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) systems¹. 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

---

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{sub}DS{-}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.

**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 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].

**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 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 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].
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. Valkチンov 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-prescribed-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.
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 [102–04]. 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-MedNLP², 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 “nivolumab” 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


100. Yada S, Aramaki E, Tanaka R, Cheng F, Kurohashi
99. Campillos L, Deléger L, Grouin C, Hamon T,
96. Yada S, Joh A, Tanaka R, Cheng F, Aramaki E,
95. Wang J, Lavender M, Hoque E, Brophy P, Kautz
91. Zhang L, Negrinho R, Ghosh A, Jagannathan
88. Gopinath D, Agrawal M, Murray L, Horng S,
SI annOtated Text corpus (MERLOT). Language
with comprehensive semantic annotations: devel
of a large clinical entity corpus. In: Proceedings
ly annotated corpus of clinical texts. J Biomed
In: Proceedings of the 12th Language Resources
knowledge: starting from critical lung diseases.
17;4(1):00ab003.
113. Rouzfarakh M, Deldar K, Froutan R, Ahmad
108. Wei CH, Peng Y, Leaman R, Davis AP, Mat
notation in the clinical domain. Transactions of the association for computational linguistics
2014:2:143-54.
Jan;42(1):99-111.
106. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical ra-
diology reports. J Am Med Inform Assoc 2007
Dec;132:103978.
114. Almusharraf F, Rose J, Selby P. Engaging Un-
118. Głowacki EM, Bernhardt JM, McGlone MS. Tai-
lored texts: An application of regulatory fit to text messages designed to reduce high-risk drinking. Health Informatics J 2020 Sep;26(3):1742763.
115. Yonek JC, Meacham MC, Ramo D, Delucchi K, Tolou-Shams M, Satte DD. The Relationship of E-Cigarette Use to Tobacco Use Outcomes Among Young Adults Who Smoke and Use Alco-
Dec;132:103978.
Dec;132:103978.

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, RuBERTa = Robustly Optimized BERT Pretraining Approach, SVM = Support Vector Machine, T5 = Text-to-Text Transfer Transformer

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

* Papers published in 2020 or later.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Objectives</th>
<th>Study design</th>
<th>Target language</th>
<th>Corpus</th>
<th>NLP task</th>
<th>Method</th>
<th>Performance [95% confidence interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valtchinov et al., 2020 [52]*</td>
<td>Identification of implantable device posing MRI safety risks</td>
<td>Retrospective study</td>
<td>English</td>
<td>Radiology reports and other clinical notes</td>
<td>NER</td>
<td>Existing tool</td>
<td>Accuracy 0.83–0.91</td>
</tr>
<tr>
<td>Trivedi et al., 2018 [53]</td>
<td>Determination to use or not to use contrast enhancement for MRI imaging</td>
<td>Retrospective study</td>
<td>English</td>
<td>Free-text MRI protocols and indications</td>
<td>Document classification</td>
<td>Existing tool</td>
<td>Accuracy 0.832</td>
</tr>
<tr>
<td>Chillakuru et al., 2021 [54]*</td>
<td>Determination to use or not to use contrast enhancement for MRI imaging</td>
<td>Retrospective study</td>
<td>English</td>
<td>Free-text MRI protocols and indications</td>
<td>Document classification</td>
<td>TF-IDF vectorization + GBDT, Word embedding + shallow NN</td>
<td>Accuracy 0.8338–0.8543</td>
</tr>
<tr>
<td>Kim et al., 2020 [57]*</td>
<td>Information extraction from pathology reports</td>
<td>Retrospective study</td>
<td>English</td>
<td>Pathology reports</td>
<td>NER</td>
<td>BERT</td>
<td>Accuracy 0.9795–0.9839</td>
</tr>
<tr>
<td>Odisho et al., 2020 [61]*</td>
<td>Information extraction of histological characteristics of prostate cancer</td>
<td>Retrospective study</td>
<td>English</td>
<td>Pathology reports</td>
<td>Document classification and token classification</td>
<td>CNN, Random forest</td>
<td>Weighted F1-score 0.972 (document), Accuracy 0.930</td>
</tr>
<tr>
<td>Dui et al., 2020 [72]*</td>
<td>Automated diagnosis of multiple psychiatric diseases</td>
<td>Retrospective study</td>
<td>English and Chinese</td>
<td>Discharge summaries</td>
<td>Multilabel document classification</td>
<td>RoBERTa</td>
<td>Micro F1-score 0.584</td>
</tr>
<tr>
<td>Moon et al., 2022 [80]*</td>
<td>Recognition of gynecological surgical history</td>
<td>Retrospective study</td>
<td>English</td>
<td>Clinical notes</td>
<td>Multiclass classification</td>
<td>NER + Rule-based approach</td>
<td>Weighted F1-score 0.76</td>
</tr>
<tr>
<td>Sterdix et al., 2020 [81]*</td>
<td>Birth risk estimation</td>
<td>Retrospective study</td>
<td>English and Dutch</td>
<td>Clinical notes and structured data</td>
<td>Binary classification</td>
<td>GBDT</td>
<td>F1-score ≥ 0.80</td>
</tr>
<tr>
<td>Barber et al., 2021 [82]*</td>
<td>Prediction of 30-day readmission after ovarian cancer surgery</td>
<td>Retrospective study</td>
<td>English</td>
<td>Radiology reports</td>
<td>Binary classification</td>
<td>(Not applicable)</td>
<td>ROC-AUC 0.70 [0.68–0.73]</td>
</tr>
<tr>
<td>Gaskin et al., 2016 [85]</td>
<td>Surveillance of postoperative complications of cataract surgery</td>
<td>Retrospective study</td>
<td>English</td>
<td>EHR notes and structured data</td>
<td>Binary classification</td>
<td>Random forest</td>
<td>ROC-AUC 0.62–0.84</td>
</tr>
<tr>
<td>Moon et al., 2018 [87]</td>
<td>Synonym replacement for better readability</td>
<td>Retrospective study</td>
<td>Finnish</td>
<td>Progress notes and nursing notes</td>
<td>Abbreviation Resolution</td>
<td>Word embedding + cosine similarity</td>
<td>Top-1 accuracy 0.3464</td>
</tr>
<tr>
<td>Gopinath et al., 2020 [88]*</td>
<td>Auto-completion of progress notes</td>
<td>Retrospective study</td>
<td>English</td>
<td>Progress notes and structured data</td>
<td>Recommendation</td>
<td>Rule-based entity type detection + Shallow NN or Bayes statistics</td>
<td>Keystroke reduction 67%</td>
</tr>
<tr>
<td>Moon et al., 2020 [89]*</td>
<td>Auto-structureization of nursing notes</td>
<td>Retrospective study</td>
<td>Finnish</td>
<td>Nursing notes</td>
<td>Sentence classification</td>
<td>LSTM</td>
<td>Accuracy 0.71</td>
</tr>
<tr>
<td>Krishna et al., 2021 [90]*</td>
<td>Automatic summarization of doctor-patient conversation</td>
<td>Retrospective study</td>
<td>English</td>
<td>Outpatient transcription</td>
<td>Summarization</td>
<td>BERT-LSTM classifier + T5</td>
<td>ROUGE-L 1.0.3838</td>
</tr>
<tr>
<td>Zhang et al., 2021 [91]*</td>
<td>Automatic summarization of doctor-patient conversation</td>
<td>Retrospective study</td>
<td>English</td>
<td>Outpatient transcription</td>
<td>Summarization</td>
<td>BART</td>
<td>ROUGE-L 1.0.3412</td>
</tr>
</tbody>
</table>

* Papers published in 2020 or later.