BMC Medical Informatics and Decision Making

RESEARCH

Open Access

From admission to discharge: a systematic review of clinical natural language processing along the patient journey

Katrin Klug[1*], Katharina Beckh[1], Dario Antweiler[1], Nilesh Chakraborty[1], Giulia Baldini[2,3], Katharina Laue[4], René Hosch[2,3], Felix Nensa[2,3], Martin Schuler[4] and Sven Giesselbach[1]

Abstract

Background Medical text, as part of an electronic health record, is an essential information source in healthcare. Although natural language processing (NLP) techniques for medical text are developing fast, successful transfer into clinical practice has been rare. Especially the hospital domain offers great potential while facing several challenges including many documents per patient, multiple departments and complex interrelated processes.

Methods In this work, we survey relevant literature to identify and classify approaches which exploit NLP in the clinical context. Our contribution involves a systematic mapping of related research onto a prototypical patient journey in the hospital, along which medical documents are created, processed and consumed by hospital staff and patients themselves. Specifically, we reviewed which dataset types, dataset languages, model architectures and tasks are researched in current clinical NLP research. Additionally, we extract and analyze major obstacles during development and implementation. We discuss options to address them and argue for a focus on bias mitigation and model explainability.

Results While a patient's hospital journey produces a significant amount of structured and unstructured documents, certain steps and documents receive more research attention than others. Diagnosis, Admission and Discharge are clinical patient steps that are researched often across the surveyed paper. In contrast, our findings reveal significant under-researched areas such as Treatment, Billing, After Care, and Smart Home. Leveraging NLP in these stages can greatly enhance clinical decision-making and patient outcomes. Additionally, clinical NLP models are mostly based on radiology reports, discharge letters and admission notes, even though we have shown that many other documents are produced throughout the patient journey. There is a significant opportunity in analyzing a wider range of medical documents produced throughout the patient journey to improve the applicability and impact of NLP in healthcare.

Conclusions Our findings suggest that there is a significant opportunity to leverage NLP approaches to advance clinical decision-making systems, as there remains a considerable understudied potential for the analysis of patient journey data.

Keywords Clinical natural language processing, Patient journey, Out-of-distribution generalization, Explainable ML, Bias

*Correspondence:
Katrin Klug
katrin.klug@iais.fraunhofer.de
Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

Natural language processing (NLP) has achieved significant success in applications such as translation, speech recognition, text generation, virtual assistants, and chatbots [1, 2]. These applications cover industrial, creative as well as lifestyle domains, and more recently, also the healthcare sector [3, 4]. Due to an increasing number of patients, rising costs and larger amounts of data, there is a high demand for automated processing of health-related documents. Hospitals struggle to provide high-quality care due to the complexity of patient histories and the high volume of medical documents generated during hospital stays, including reports from pathology, radiology, laboratory, surgery, and care documentation [5]. This information and is crucial for any decision on diagnostics, therapy or subsequent care. Significant effort is dedicated to the tasks of writing, filing, sorting, searching, retrieving, issuing, and managing medical records by the clinicians. But it is nearly impossible for clinicians to process this bulk of information [5]. Therefore, it is highly desirable to supply healthcare professionals as well as patients with information contained in these full texts by extracting data, mapping it onto clinical guidelines or otherwise inform their decisions. Hence, almost all Clinical Decision Support Systems (CDSS) depend on a continuous and reliable processing of clinical text [6]. Despite the promising capabilities of NLP for enhancing clinical decision-making and operational efficiency, its integration into real-world healthcare settings remains limited due to challenges such as data quality, lack of standardization, and inadequate alignment with clinical workflows [7]. This study aims to address these challenges and provide solutions to facilitate the integration of NLP in clinical environments. The significance of this research lies in its potential to bridge the gap between NLP research and its application, ultimately contributing to improved patient outcomes and operational efficiency. Our goal is to equip researchers with established and successful approaches for clinical NLP. Together with the mounting number of publications in this research area, this motivates a systematic survey of existing approaches.

In this survey, we report on the current state of research in clinical NLP along the different stages of a patient's journey through a hospital. In collaboration with doctors as domain experts, we have created a prototypical patient journey. In total, we reviewed 8.527 papers, applying a filtering and screening process to include medical and clinical papers. On the one hand, we used NLP-related tags to map the papers to relevant NLP tasks, models, datasets, and data languages. On the other hand, we used clinical tags, such as general patient journey and patient journey documents, to ensure mapping the NLP applications to the actual patient journey. Previous work, such as that by [7], provides a foundation for understanding the practical considerations necessary for developing effective clinical NLP systems.

We identify gaps between research and clinical application of NLP in hospitals, as well as areas that require further exploration and development. In particular, our results show that there is a lack of research in developing trustworthy models, and we thus highlight distinct challenges in this field of NLP in the clinical setting and suggest an outline on how to address them during development.

We begin by describing related work in the area of NLP for hospital documents. The subsequent section describes in detail a prototypical patient journey, along which medical documents are created, processed and consumed by hospital staff and patients themselves. We describe our methodology of identification, selection and extraction of relevant publications in the literature and the key insights obtained. In the main section, we map recognized concepts onto our framework consisting of multiple technical and medical dimensions and follow up with an analysis and discussion of the results. The final section concludes with a description of overarching patterns and suggestions for the applications of NLP systems in clinical practice.

Related work

The use of NLP in medicine has been the focus of several surveys in recent years. Topics that have been investigated include deep learning architectures deployed in medical imaging and NLP [8], or the implementation of task-oriented dialogue systems for healthcare applications [9]. Other studies have concentrated on NLP systems for capturing and standardizing unstructured clinical information and generate structured data [10]. Most of the mentioned surveys on NLP in the medical domain focus on a specific task, such as converting image to text or dialogue systems, and do not provide a holistic view of NLP applications in healthcare.

Recently, some studies have explored the patient journey in the hospital. While [11] applied process mining techniques to the patient journey to improve the patients' satisfaction, [12] discussed general AI opportunities along the patient journey. To the best of our knowledge, no prior research has focused on mapping the patient journey onto NLP tasks in research. Therefore, in our survey, we concentrate on this mapping to analyze the current NLP research and applications along the patient journey by reviewing relevant publications. Unlike previous studies that focus on specific tasks, our review provides a holistic view of NLP applications throughout the patient journey, identifying gaps in areas such as After Care and Smart Home. Our approach integrates

NLP into various stages of the patient journey, offering a detailed perspective that previous studies lack. By mapping the patient journey onto NLP tasks, we provide insights into how NLP can be utilized not only in clinical settings but also in post-discharge and home care scenarios. This broadens the scope of NLP applications beyond traditional settings. Furthermore, our approach identifies overlooked areas, offering a roadmap for future research and development in NLP applications across patient care.

Patient journey

To better illustrate the amount of unstructured documents that patients encounter during their hospital stay, we employed a case study approach to present the hospital journey of a cancer patient. Specifically, we focused on the patient journey of a lung cancer patient, as it is one of the most commonly diagnosed subtypes of cancer, and cancer is the second leading cause of death in the western world [13].

The patient journey begins with a suspected diagnosis of lung cancer, followed by complex diagnostic procedures and resulting in cancer treatment [14, 15], as shown in Fig. 1, where we also highlighted the emerging documents during this process. Medical information systems typically document these findings and information in unstructured text, except for laboratory test results, which are usually available in a structured format.

Most commonly, lung cancer is suspected based on arising symptoms or as an incidental finding in an imaging study. In the next step, the patient gets hospitalized for further diagnostic procedures, usually in the pneumology department. This diagnostic pathway starts with the anamnesis and a physical examination by the physician as well as laboratory tests, followed by a tumor biopsy and a lymph node sampling for histological examination and staging. The tumor staging is completed by performing further imaging studies. For the staging of lung cancer, a positron emission tomography-computed tomography (PET-CT) and a brain magnetic resonance imaging (MRI) are the gold standard. To evaluate the cardiopulmonary function of the patient, functional tests by means of electrocardiogram (ECG), transthoracic echocardiogram (TTE) and pulmonary function tests are carried out. Each of these steps produces one or multiple reports. When all the diagnostic information is available, the treatment strategies are discussed in a multidisciplinary lung cancer tumor board consisting of medical oncologists, radiation oncologists, pulmonologists, thoracic surgeons, radiologists and/or nuclear medicine specialists and pathologists. If a chemotherapy

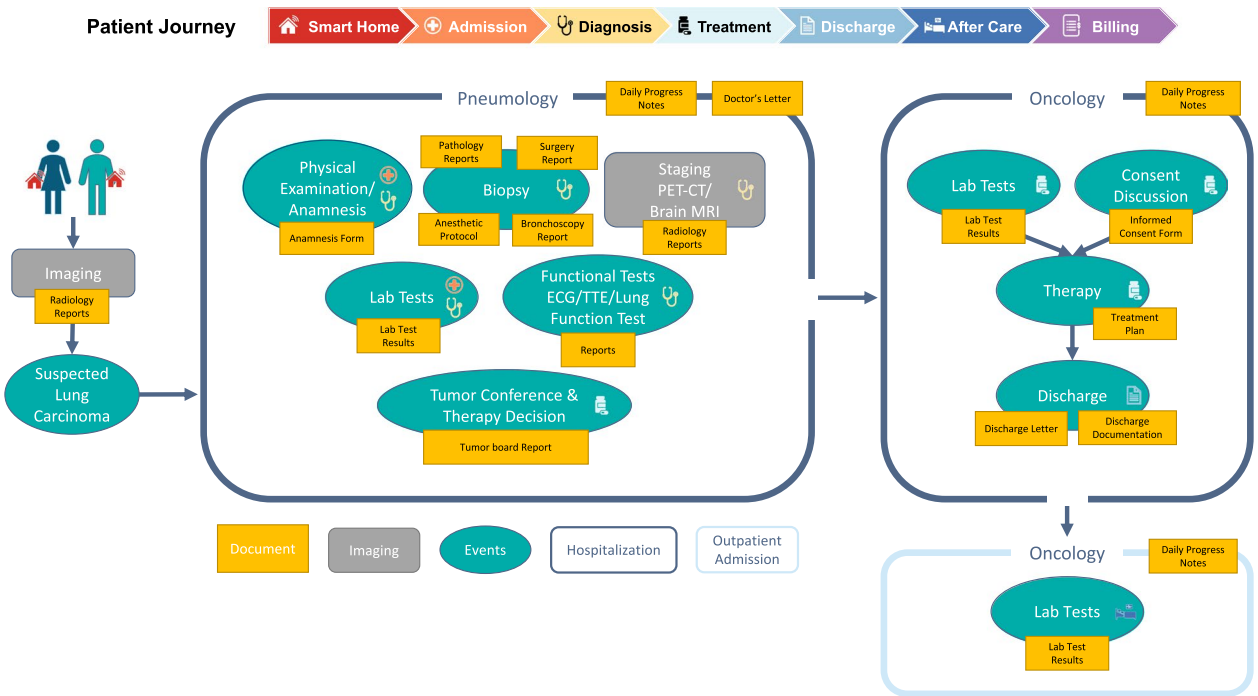

Fig. 1 Typically, a suspicion of lung carcinoma leads to admission to the pneumology department. Multiple tests are conducted to reach a diagnosis. The treatment options for the patient are discussed by a multidisciplinary tumor board and the patient is transferred to the oncology department to undergo the chosen therapy. Once completed, the patient is discharged from the hospital, but may continue to visit for follow-up checks to ensure effective treatment. Documents collected during this journey are highlighted in yellow. The steps in this example are marked with the symbols of the corresponding phases of the general patient journey, shown at the top

is recommended, the patient is transferred to the oncology department. Once the consent discussion has been completed, a systemic anticancer therapy such as a chemotherapy and/or an immunotherapy is applied. Following the systemic therapy, the patient gets discharged. Approximately one week after the application of the therapy, an ambulant laboratory test is recommended. If necessary, the patient returns for the second cycle of chemotherapy after a typical waiting period of two to three weeks.

From this use case we initially identified five main patient journey stages: **Admission**, **Diagnosis**, **Treatment**, **Discharge** and **After Care**. In each of these stages, different documents are collected, processed and used by other clinicians in later stages. As shown in Fig. 1, the corresponding events for each stage have been mapped using respective symbols. The diagnostic phase is typically the most document-intensive, as patients undergo numerous tests and procedures to obtain a suitable diagnosis. The treatment phase also generates a significant amount of documentation, owing to the close monitoring of the patient's progress to ensure that the treatment is proceeding smoothly. Our particular patient journey involves fewer documents in the other three stages. Although not directly evident in the patient journey, we have included the initial stage of the journey, **Smart Home**, as Internet of Things (IoT) applications are becoming increasingly relevant in the healthcare sector [16]. Patients may, for example, bring heart rate measurements monitored using their smartwatches, which could be used as an additional diagnostic tool. Another part of the journey that does not directly influence the patient care is **Billing**, which is a source of multiple unstructured documents.

In summary, the hospital journey of a patient, in this example a cancer patient, generates a significant amount of structured and unstructured documents. To better understand this process, we have divided the journey into seven main stages, where each stage produces different types of documents that are crucial to the overall care of the patient. By recognizing the document-intensive nature of the patient journey and the potential for unstructured data to impede care, we can begin to explore the benefits of implementing NLP technologies to streamline document handling and improve patient outcomes.

Methods

In order to analyze the transfer of NLP research into the clinical domain and map the actual use of NLP throughout the patient journey, we conducted a systematic review of 8.527 papers based on publication venue, date, and title combined with a keyword search as our selection criteria. The tagging was performed for two dimensions. The first dimension concentrated on NLP-related tags to map the papers to relevant NLP tasks, models, datasets, and data languages. The second dimension focused on clinical tags, such as general patient journey and patient journey documents. The final list of publications was then screened with NLP-related and patient journey related tags. A team of four reviewers annotated the papers, and the papers were equally split among the reviewers. Each paper was annotated by two reviewers and in case of doubts, a third reviewer was used for tie-breaks. In visualizing the results, we employed Python along with its packages including Seaborn, Matplotlib, Pandas, Plotly, and Sankey, ensuring comprehensive data representation. In the following, we provide an overview of the methodology used in our review.

Search criteria and screening process

In the following, we describe our search criteria and screening process for selecting literature.

Publication venue. In our systematic review we focused on articles published in NLP conferences from the ACL anthology (ACL, EMNLP, COLING, CoNLL, EACL, NAACL, AACL) and workshops from end of 2018 to December 2022. Specifically, we targeted workshops that have a medical research focus, like BioNLP, NLPMC, SMM4H, ClinicalNLP, LOUHI. All articles were last extracted in January 2023.

Title screening. To further refine the search, we employed a keyword filtering process. We selected relevant keywords through discussions with healthcare professionals in the clinical domain and screened the titles of the initial list of papers. The following list of keywords was used: *medical, medicine, health, care, patient, treat, cancer, hospital, surgery, surgical, drug, emergency, doctor, surgeon, human, disease, diagnosis, trauma, report, discharge, clinical.*

Abstract and paper screening.

Relevance and Medical Domain: Next, we filtered our remaining paper list by screening the abstracts and excluding papers that are not relevant for the medical domain. Additionally, we excluded papers that were research or tutorial proposals, or demo papers.

Clinical Screening: As our research focuses on clinical NLP and the patient journey in a hospital, we further refined the list by identifying the papers that are relevant for the clinical domain. Research analyzing bio markers or social media posts were excluded by our clinical screening process. The initial collection, based on the selection of the publication venue and the years, consisted of 8.527 papers. The filtering process led to 609 publications after the keyword search in our title

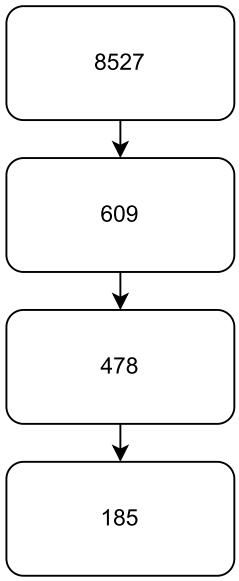

Fig. 2 Amount of papers per screening process step for the selection of the reviewed paper list

Table 1 Patient journey tags

Category	Tags
General patient journey	P1: Smart Home (e.g. preclinical data, home-monitoring devices)
	P2: Admission
	P3: Diagnosis
	P4: Treatment
	P5: After Care
	P6: Discharge
	P7: Billing
	P8: Other
Patient journey report type	R0: Admission notes
	R1: Radiology report
	R2: Discharge letter
	R3: Documented histories
	R4: Pathology report
	R5: Tumour conference decisions
	R6: Lab results
	R7: Surgery report
	R8: Other

screening, 478 after relevance and medical domain filtering and remaining 185 clinical domain papers (see Fig. 2).

Tagging process

Our review involved mapping every paper of our screening process to several NLP-related tags, with the aim of identifying which models, tasks, datasets, and data languages are most commonly used in healthcare NLP research. To identify the current applications of NLP research in the hospital, we included tags for patient journey and document types. Specifically, we assigned tags to each paper based on the stage of the patient journey that was being addressed (e.g., diagnosis, treatment, admission), as well as the type of patient journey document that was being analyzed (e.g., clinical notes, discharge summaries, radiology reports) (see "Patient journey" section). For this part of the analysis, we only focused on the papers left after the Clinical Screening process, as our patient journey concentrates on a hospital patient (see "Search criteria and screening process" section). We assigned multiple tags where applicable, e.g. when multiple datasets were used. Our detailed tagging scheme can be seen in Tables 1 and 2.

Results

In this section, we present our findings in relation to (1) NLP systems in the healthcare domain and (2) along the patient journey.

Mapping NLP tags

In the following, we describe the results of our review with respect to the NLP methodologies and datasets implemented in the healthcare domain.

Dataset language: Various studies have analyzed or explored datasets consisting of multiple data languages. Through the analysis of 487 papers, we observed that English was the most frequently used dataset language (419). The second and third most used dataset language were Spanish (36) and Chinese (25). The remaining 237 languages were classified under the 'Other' category (see Fig. 3).

Dataset type: In terms of datasets, we found that patient related data, like electronic health records, were the most commonly used sources of data (27%), followed by clinical studies (20.7%), and forum posts, chat logs, social media datasets (19.1%), as demonstrated in Fig. 3.

Model type: Figure 3 displays that transformer-based models were the most commonly used type of NLP model across a variety of tasks (44.94%), followed by recurrent neural networks (RNN) (20.39%). As shown in Fig. 4, in 2019, RNNs were still used more frequently than transformer-based models. The use of transformer-based models increased over a four-year period, culminating in a peak in 2021 and 2022.

NLP Task: Finally, we observed that certain tasks, such as classification with almost 30%, information extraction with 26.81% and text generation/text summarization which account for 12.52%, were more frequently studied than others.

Table 2 General NLP-related tags

Category	Tags
Data type	D1: All patient related records
	D2: Clinical studies
	D3: Registry data
	D4: Protein data
	D5: Genome data
	D6: Forum posts, chatlogs, social media
	D7: Speech data, dialogue data
	D8: Image data
	D9: Knowledge graph, thesaurus
	D10: Medical online information (Wikipedia, drug information, FAQs, etc.)
	D11: Patents
	D12: News articles and press releases
	D13: Clinical guidelines
Data language	free text, e.g. English, German
Task	T1: Classification
	T2: Information extraction
	T3: Clustering
	T4: Text generation
	T5: Embeddings/representations
	T6: New dataset creation
	T7: Question answering
	T8: Text summarization
	T9: Translation
	T10: Reinforcement learning
	T11: Recommender system
	T12: Natural Language Inference and entailment
	T13: Topic model
	T14: Probing
	T15: Ranking
Secondary task	S1: Explainability
	S2: Domain adaptation
	S3: Bias, fairness
	S4: Resource-awareness
Model type	M1: Transformer-variants (BERT, RoBERTa etc.)
	M2: Convolutional Neural Nets (CNNs)
	M3: Recurrent Neural Nets (RNN, LSTM)
	M4: Statistical models (Bayes, conditional probabilities, CRF)
	M5: Graph Neural Networks (GNNs)
	M6: Dimension reduction
	M7: Graphical models (PGM)
	M8: Generative Adversarial Networks (GANs)
	M9: Rule-based models
	M10: Decision trees, Random Forest
	M11: Support Vector Machines (SVM)
	M12: K-nearest neighbors (kNN)
	M13: Pointer generator model
	M14: Feedforward neural network
	M15: Logistic regression
	M16: Linear regression
	No contribution

Mapping of patient journey

Analyzing the clinical patient journey, we observe that most of the clinical NLP papers focus on applications during the Diagnosis, Admission and Discharge phase of the patient, while referring to admission notes, radiology reports and discharge letters. It is remarkable that the most researched patient journey step is the Diagnosis while taking into account mostly radiology reports. As shown in Fig. 5, paper with the focus on the Treatment of the patients do not use a specific document type as a focal point, but an evenly distribution of admission notes, radiology reports, documented histories, discharge summaries and other document types. In contrast to that, patient journey steps like Smart Home, After Care, or Billing are less represented in the clinical NLP literature.

Discussion

We observe that most of the publications in the medical NLP literature use English datasets, see also [17]. This indicates that other languages are under-researched in the medical domain whereby potential of clinical NLP application gets lost. Focusing on English data leads to an imbalance between non-English and English medical applications [18]. NLP models that are trained solely on English data may not perform as well when applied to other languages [19], because language models often rely on patterns and structures that are specific to a particular language, and these patterns may not be present in other languages [20]. Furthermore, by expanding the scope of research to other languages, researchers can uncover new patterns and structures that may not be present in English, leading to new breakthroughs and advancements in the field [19]. We already observe attempts to include non-English datasets. For example, most studies that dealt with Spanish datasets were published in the sixth and seventh Workshop on Social Media Mining for Health Applications and assessed Spanish tweets regarding health conditions [21, 22] or fifth Workshop on BioNLP Open Shared Tasks [23].

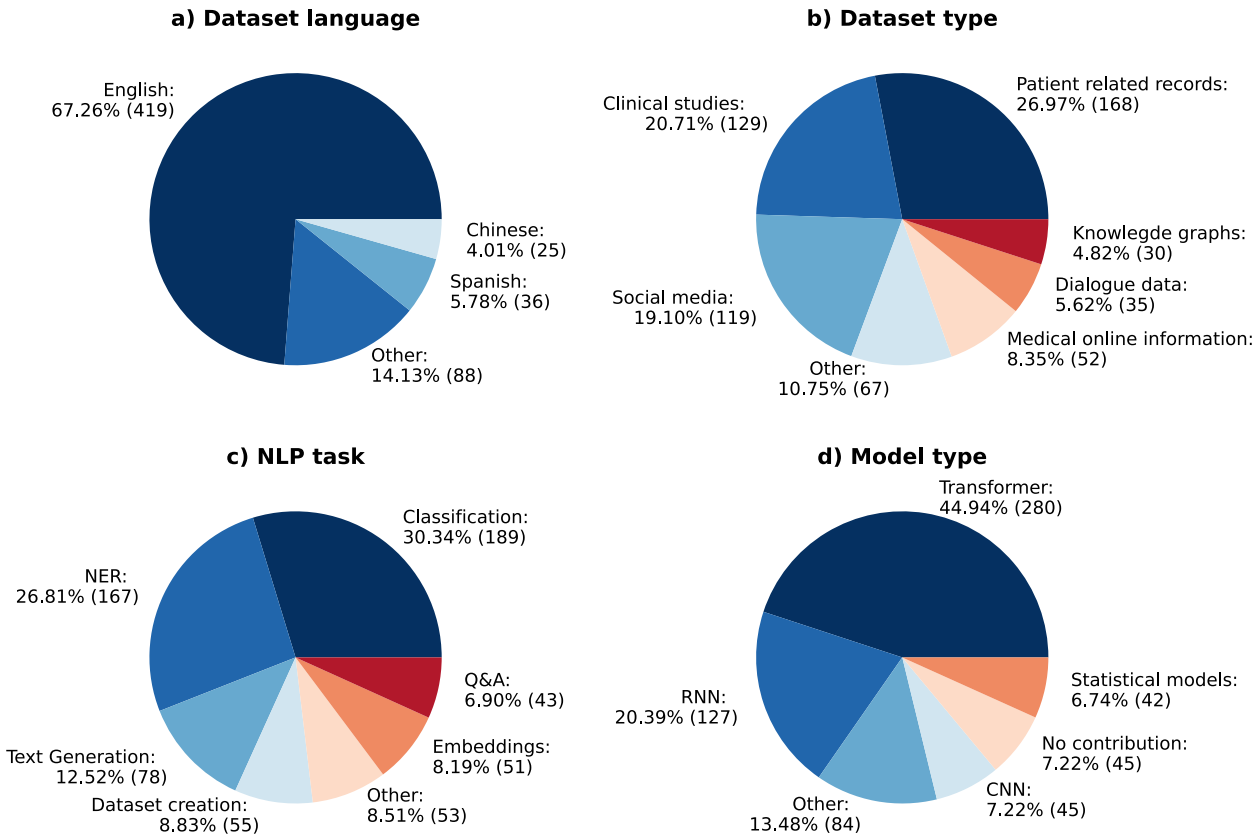

Fig. 3 Distribution of the number of papers per NLP related tag category: (**a**) dataset language (**b**) dataset type (**c**) NLP task (**d**) model type.Values that fell below the 5% threshold were aggregated into "Other" category for the purposes of analysis, except for dataset language, where we display the top three dataset languages

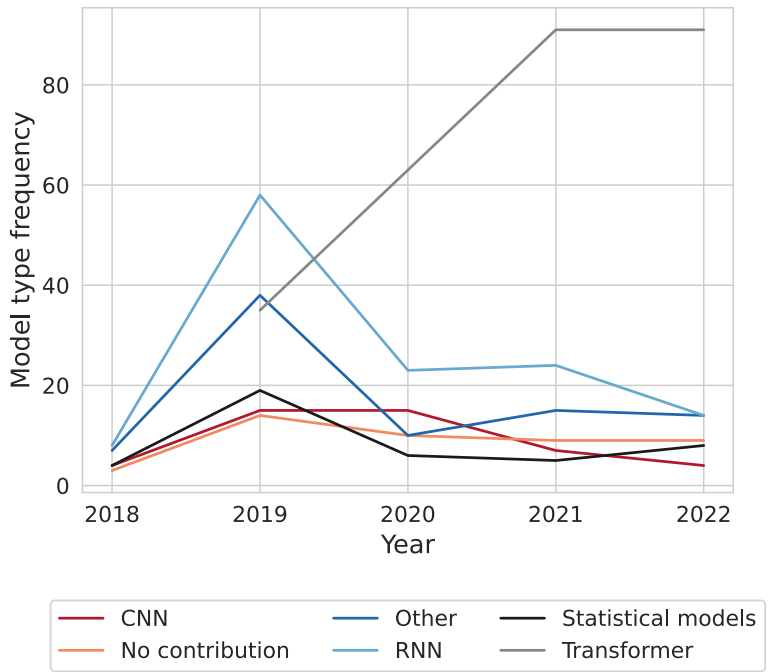

Fig. 4 Development of model types used in NLP research over the past years

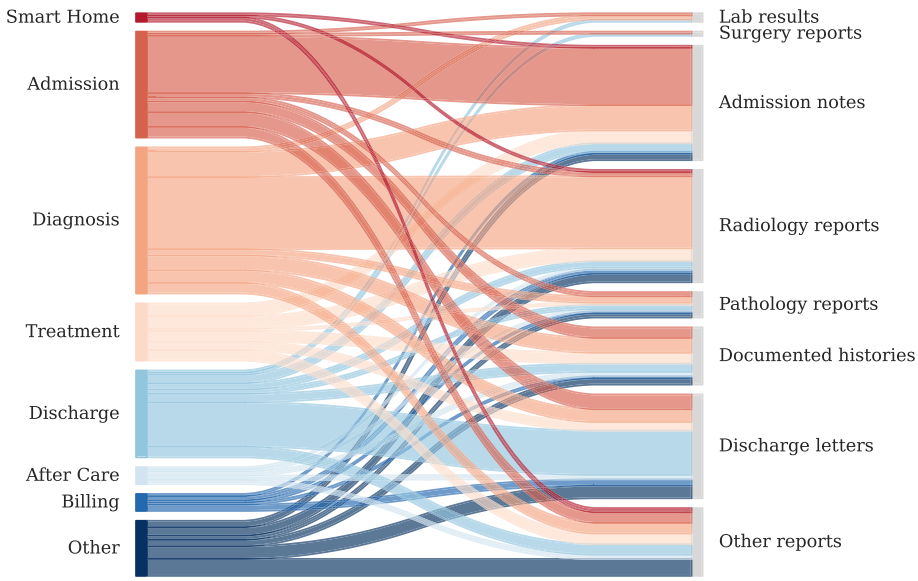

Fig. 5 Patient Journey results: Comparison of patient journey steps (left side) with the patient journey documents (right side). The width of each stream shows how often the patient journey step or document type appeared in the reviewed papers

Looking at the data types, it is noteworthy that patient related records are used most frequently. Wornow et al. [17] found that there is an over-representation of models that were trained on the MIMIC-III dataset, as it is one of the few public available patient related datasets. Other publicly available datasets are needed to create models that are trained on larger clinical data with current knowledge about diseases and treatments [17]. Our findings show that current NLP research tends to focus on specific types of documents, such as radiology reports, discharge letters, and admission notes. There is a significant opportunity in analyzing a wider range of medical documents produced throughout the patient journey, such as care and disease progression documentation. Expanding the scope of analyzed documents to include a more diverse range of patient data will enhance the applicability and impact of NLP in healthcare. It is striking that although transformer models have been discussed in research since 2017 when they were first invented [24], they were mainly used in medical applications from 2020 onwards. This could indicate that there is a general delay in applying novel methods in the medical domain. While transformer models have shown promising results in research papers after 2017, implementing them in real-world applications may be challenging. Firstly, transformer models are often large and computationally intensive, which can make them difficult to run on resource-limited devices [25]. Additionally, transformer models require large amounts of high-quality training data. In the healthcare domain, obtaining such data can be difficult and limited due to sensitive patient related data that needs to be anonymized first [26–28]. Furthermore, transformer models are often referred to as "black boxes" because it can be difficult to understand how and why they predict a specific output. This lack of interpretability can make it challenging to use these models in the medical domain, where transparency and accountability are important [29–31].

Remarkable is the extent to which clinical systems can be supported by NLP technologies. We have shown that a patient generates a significant amount of structured and unstructured documents throughout the journey in a hospital. We observed that specific steps of the patient journey are researched more often than others. While Diagnosis and Admission are areas that are researched primarily in the clinical NLP community, there seem to be patient journey steps that are under-researched, e.g. Smart Home, Treatment, After Care, Billing, even though a lot of documents are produced for every patient in the hospital (see "Patient journey" section). In terms of documents, admission notes, radiology reports and discharge letters are used most frequently, which is in line with the previously analyzed patient journey steps (see Fig. 5). Patient journey steps such as Admission and Diagnosis are often considered to be critical in the patient journey, where early detection and intervention can have a significant impact on patient outcomes [32]. This may make them a priority for research and development of NLP models. One reason could again be that researchers may focus on points of the patient journey where high-quality data is available like admission notes, radiology and pathology reports. We observe that radiology reports are mainly used for the

diagnosis in our review, which indicates that there is still a huge potential of clinical NLP technologies analyzing other report types than radiology reports for improving the diagnosis of a patient, as shown in Fig. 1. One reason could be that radiology and imaging reports are one part of the MIMIC-III dataset and predominantly used for research. Additionally, some data that is available and not related to patient data may be medically related but not clinically related. Examples are investigation of social media data to analyze the symptoms of COVID patients [33], detect patients' emotional states [34] and mental illnesses [35, 36] or identify adverse drug reactions [37–40]. This type of data provides valuable insights into health trends and biological mechanisms, contributing to the broader understanding of medical science. However, it may not have immediate implications for patient care, unlike clinically related data, which includes patient history, diagnostic test results, and treatment outcomes. Our review shows that there is still a huge potential to support clinical decision systems with NLP methodologies, as the application opportunities lack behind the application reality. Researchers should explore NLP applications in Treatment and Billing phases to automate routine tasks, thereby reducing administrative burden and enhancing patient care which can lead to more accurate diagnoses and effective treatment plans. Practitioners can benefit from implementing NLP tools for better patient monitoring and follow-up in After Care. Furthermore, interdisciplinary collaboration between NLP researchers, clinicians, and healthcare administrators is crucial. Such collaborations ensure that NLP innovations are both technically sound and practically useful in clinical settings. Additionally, developing user-friendly NLP applications that are intuitive and easy to use can facilitate quicker adoption into clinical practice. By focusing on these aspects, both medical practitioners and researchers can use NLP methods to improve patient outcomes, streamline clinical workflows, and improve medical research.

Challenges

There are several challenges which might prevent or slow down the process of applying NLP technologies in the hospital setting. While primary down-stream tasks can now be reasonably tackled, we are especially facing challenges in the field of trustworthiness. Contrary to our expectation, the reviewed papers largely omit this topic (ca. 16% of papers address trustworthy ML topics). In the following section we address this research opportunity and concentrate on the discussion of three challenges in the field: out-of-distribution generalization, explainability and bias.

Out-of-distribution generalization

One of the fundamental assumptions in supervised machine learning is the existence of identical and independently distributed data. Models perform well provided that test-time data points are distributed similarly to those used for training. In practice, we may have several sources of distribution shift between the training environment and the setting in which the model is deployed, leading to a lack of performance.

One of the sources of distribution shifts is subpopulation shift [41]. The training dataset may consist of data points that have the same label, while simultaneously having multiple distinct *subgroups* among them, i.e. the label only coarsely describes the meaningful variation within the population. A data subgroup might contain spurious correlations between its features and labels that do not hold outside this subgroup. If such subgroups are large enough, a model trained by minimizing empirical risk will latch onto these spurious correlations and underperform on "minority subgroups". Shim et al. [42] investigate imbalance in a medical code prediction dataset in terms of demographic variables, and observe the issue of subpopulation shift while analyzing the performance differences of the model across demographic groups. This problem may be exacerbated when the model is tested in a deployment scenario with different distributions of demographic groups than that encountered during training. Holderness et al. [43] show that off-the-shelf sentiment classification models trained on general domain data do not perform very well on psychiatric patient health records. They further demonstrate that domain adaptation methods based on self-training and k-nearest neighbors can be used to adapt off-the-shelf models by leveraging a corpus of unlabeled electronic health record data.

Medical records which are written by clinicians from different specialties usually differ in terms of writing styles or terminologies used. In order to train Named Entity Recognition (NER) models on medical records, human-annotated datasets are needed. But the cost of human annotation makes it difficult to create labelled datasets in all specialties. Wang et al. [44] propose a label-aware domain transfer method for medical NER that learns a close feature mapping between source and target domains. This enables NER models trained on one specialty to be conveniently applied to another one with minimal annotation effort. Liu et al. [45] uses domain-adversarial training to learn whether a pair of disease phrases from different domains are semantically similar without requiring a lot of pairwise labelled data.

Explainable machine learning

One key challenge in adopting machine learning systems in the clinical domain is missing transparency [29–31]. NLP systems, in particular, suffer from opaqueness due to a reliance on deep neural networks. This is evident in the results of the literature analysis: Over 70% of papers rely on transformer variants, CNNs or RNNs, which are notoriously hard to interpret.

The field of explainable machine learning offers methods to address the lack of transparency [46–50]. Explainability in the clinical context is relevant for compliance legislation, system improvements and verification [51]. Explanations have different forms, such as text highlighting, rules or examples. The most prevalent explanation form in NLP is feature attribution, which typically highlights the features, e.g. tokens, that contribute most to a prediction [52]. While abundant explanation methods are available for predictive tasks, explanations for generative tasks are lacking and present an open research topic.

From the reviewed work, 21 papers (4%) explicitly mention explainability or interpretability. As is common in the NLP domain, the terms are mostly used interchangeably [53]. Roughly 40% use feature attribution as explanation form, which is in line with other reviews [52]. Another 40% integrate interpretable components or can be considered interpretable-by-design. Six papers report quantitative or qualitative evaluation, incl. three works which evaluate with one or two clinicians. In contrast, the majority of papers claims that the model is more interpretable or explainable without any quantification. Anecdotal evidence is common and a fundamental flaw in the field [54].

Ghassemi et al. [55] argue that current explainability methods are not sufficient for certain purposes in the clinical domain and argue to focus more strongly on validation practices. The explanation purpose is often not defined, which hinders the assessment of usefulness. In addition, we agree that rigorous validation is important and we start to see works in this direction, e.g. [56]. However, we emphasize that e.g. the robustness field is facing similar challenges with guarantees. A sole focus on validation is not sufficient to tackle transparency requirements. For this reason, we call for purpose-driven development and adequate evaluation to derive in which ways explanations are most beneficial for the clinical context.

Bias

In the medical domain, data bias is prevalent and imminent. While biomedical publications are mainly affected by reporting bias [57], medical record datasets can contain bias from multiple sources, including authorship, target audience, local practices, type of trigger, available time, deployed software or monetary incentives [58]. Whether employed dataset(s) are representative for a patient population is heavily dependent on data collection practices. For instance, in the case of acute kidney injury (AKI), less than 29% of all clinically identified AKI patients receive a corresponding ICD code in their patient record [59]. NLP models have trouble to differentiate sentences describing normalities from important abnormalities in radiology reports [60]. For machine learning systems that support clinical staff and patients in taking informed decisions, non-discrimination of protected groups is an essential goal. Handling bias in machine learning often consists of detecting and, when indicated, reducing bias. Detection is often driven by calculating statistical fairness metrics, such as *Group fairness* or *Equalized Odds*. It must be emphasized that no single metric is sufficient on its own, instead each application requires a combination of metrics, selected by a careful consideration of moral reasoning and domain-specific challenges [61]. Debiasing word embeddings and post-processing via equalized-odds can improve downstream clinical NLP tasks [62].

Limitations

While we believe that our selection of NLP conferences provides valuable insights into current trends and advancements in the field, it is important to acknowledge the limitations of our methodology. Specifically, we chose to focus solely on NLP conferences from the ACL anthology and did not include general ML conferences or application-focused conferences from the medical domain in our analysis. This decision was made in order to provide a more focused and in-depth analysis of the technical aspects of the field. Future research may benefit from including a wider range of different NLP-related conferences and medical-related conferences in the analysis to better understand the intersection of technical advancements and real-world applications. While our study primarily focuses on mapping the patient journey onto NLP tasks, future work should expand on potential approaches to address bias mitigation and enhance model explainability. Addressing these challenges will further strengthen the deployment of NLP in healthcare, ensuring that the systems are fair, transparent and trustworthy.

Conclusion

In this paper, we conducted a systematic literature review and mapped clinical NLP research onto a prototypical patient journey in the hospital. Specifically, we reviewed which dataset types, dataset languages, model architectures and tasks are researched in current clinical NLP research. Our results show that, while a patient's hospital journey produces a significant amount of structured and unstructured documents, certain steps and documents

receive more research attention than others. Diagnosis, Admission and Discharge are clinical patient steps that are researched often across the surveyed paper. In contrast, we found that Treatment, Billing, After Care, and Smart Home are under-researched. Additionally, clinical NLP models are mostly based on radiology reports, discharge letters and admission notes, even though we have shown that many other documents are produced throughout the patient journey. Our findings suggest that there is a significant opportunity to leverage NLP approaches to advance clinical decision-making systems, as there remains a considerable understudied potential for the analysis of patient journey data.

Abbreviations

NLP	Natural Language Processing
CDSS	Clinical Decision Support Systems
PET-CT	Positron Emission Tomography-Computed Tomography
MRI	Magnetic Resonance Imaging
ECG	Electrocardiogram
TTE	Transthoracic Echocardiogram
IOT	Internet of Things
NER	Named Entity Recognition
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
AKI	Acute Kidney Injury

Acknowledgements
We would also like to thank Johann Jasper Schulze Buschhoff for his support in data provision and annotation.

Authors' contributions
SG, DA contributed to the original idea and design of the study. KK, KB, NC and DA conducted the literature review and annotated the dataset. KK collected and analyzed the data. KK, KB, GB, KL, NC and DA co-wrote the manuscript. SG, RH, FN and MS reviewed and improved the manuscript. All authors critically revised the manuscript and approved its final content.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was done within the project SmartHospital.NRW with grant number 005-2011-0041/2 and project number 2011ki001b, funded by the Ministry for Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia, Germany.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Fraunhofer IAIS, Sankt Augustin, Germany. [2]Institute of Interventional and Diagnostic Radiology and Neuroradiology, University Hospital Essen, Essen, Germany. [3]Institute for Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany. [4]West German Cancer Centre, University Hospital Essen, Essen, Germany.

Received: 22 December 2023 Accepted: 20 August 2024
Published online: 29 August 2024

References

1. Paaß G, Giesselbach S. Foundation models for natural language processing–pre-trained language models integrating media. 2023. arXiv preprint arXiv:2302.08575.
2. Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, et al. A comprehensive survey on pretrained foundation models: a history from bert to chatgpt. 2023. arXiv preprint arXiv:2302.09419.
3. Wu H, Wang M, Wu J, Francis F, Chang YH, Shavick A, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. NPJ Digit Med. 2022;5(1). https://doi.org/10.1038/s41746-022-00730-6.
4. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–65.
5. Danladi Garba K, Yahaya I. Significance and challenges of medical records: a systematic literature review. Adv Librariansh. 2018;9:26–31.
6. Reyes-Ortiz JA, Gonzalez-Beltran BA, Gallardo-Lopez L. Clinical decision support systems: a survey of NLP-based approaches from unstructured data. In: 2015 26th International Workshop on Database and Expert Systems Applications (DEXA). IEEE; 2015. pp. 163–7. https://doi.org/10.1109/dexa.2015.47.
7. Tamang S, Humbert-Droz M, Gianfrancesco M, Izadi Z, Schmajuk G, Yazdany J. Practical Considerations for Developing Clinical Natural Language Processing Systems for Population Health Management and Measurement. JMIR Med Inform. 2023;11:e37805. https://doi.org/10.2196/37805.
8. Pandey B, Kumar Pandey D, Pratap Mishra B, Rhmann W. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. J King Saud Univ Comput Inform Sci. 2022;34(8, Part A):5083–5099. https://doi.org/10.1016/j.jksuci.2021.01.007.
9. Valizadeh M, Parde N. The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: Association for Computational Linguistics; 2022. pp. 6638–6660. https://aclanthology.org/2022.acl-long.458. Accessed 1 Jan 2023.
10. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform. 2017;73:14–29.
11. Nazir T, Mushhood Ur Rehman M, Asghar MR, Kalia JS. Artificial intelligence assisted acute patient journey. Front Artif Intell. 2022;5:962165.
12. Arias M, Rojas E, Aguirre S, Cornejo F, Munoz-Gama J, Sepúlveda M, et al. Mapping the patient's journey in healthcare through process mining. Int J Environ Res Public Health. 2020;17(18):6586.
13. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209–49.
14. Postmus PE, Kerr KM, Oudkerk M, Senan S, Waller DA, Vansteenkiste JF, et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2017;28, Suppl 4:iv1–iv21.
15. Planchard D, Popat S, Kerr KM, Novello S, Smit EF, Faivre-Finn C, et al. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2018;29, Suppl 4:iv192–iv237.
16. Abdulmalek S, Nasir A, Jabbar WA, Almuhaya MAM, Bairagi AK, Khan MAM, et al. IoT-Based Healthcare-Monitoring System towards Improving Quality of Life: A Review. Healthcare. 2022;10(10):1993.
17. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med. 2023;6(1). https://doi.org/10.1038/s41746-023-00879-8.

18. Rojas M, Dunstan J, Villena F. Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing. In: Proceedings of the 4th Clinical Natural Language Processing Workshop. Seattle: Association for Computational Linguistics; 2022. pp. 87–92. https://aclanthology.org/2022.clinicalnlp-1.9. Accessed 1 Jan 2023.
19. Pilan I, Brekke PH, Dahl FA, Gundersen T, Husby H, Nytrø Ø, et al. Classification of Syncope Cases in Norwegian Medical Records. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. pp. 79–84. https://aclanthology.org/2020.clinicalnlp-1.9. Accessed 1 Jan 2023.
20. Ehsani R, Niemi T, Khullar G, Leivo T. Clinical Data Classification using Conditional Random Fields and Neural Parsing for Morphologically Rich Languages. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis: Association for Computational Linguistics; 2019. pp. 149–155. https://aclanthology.org/W19-1919. Accessed 1 Jan 2023.
21. Magge A, Klein A, Miranda-Escalada A, Al-garadi MA, Alimova I, Miftahutdinov Z, et al., editors. Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task. Mexico City: Association for Computational Linguistics; 2021. https://aclanthology.org/2021.smm4h-1.0. Accessed 1 Jan 2023.
22. Gonzalez-Hernandez G, Weissenbacher D, editors. Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. Gyeongju: Association for Computational Linguistics; 2022. https://aclanthology.org/2022.smm4h-1.0. Accessed 1 Jan 2023.
23. Jin-Dong K, Claire N, Robert B, Louise D, editors. Proceedings of the 5th Workshop on BioNLP Open Shared Tasks. Hong Kong: Association for Computational Linguistics; 2019. https://aclanthology.org/D19-5700. Accessed 1 Jan 2023.
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. CoRR. 2017. arXiv:1706.03762.
25. Kim G, Cho K. Length-Adaptive Transformer: Train Once with Length Drop, Use Anytime with Search. 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). Online: Association for Computational Linguistics; 2021. pp. 6501–6511. https://aclanthology.org/2021.acl-long.508. Accessed 1 Jan 2023.
26. 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. Brussels: Association for Computational Linguistics; 2018. pp. 2033–2042. https://aclanthology.org/D18-1228. Accessed 1 Jan 2023.
27. Raghavan P, Patwardhan S. Question Answering on Electronic Medical Records. In: Proceedings of the 2016 Summit on Clinical Research Informatics. San Francisco: AMIA; 2016. http://knowledge.amia.org/amia-59309-cri2016-1.3011827/t004-1.3012641/f004-1.3012642/a103-1.3012719/a105-1.3012714.
28. Pampari A, Raghavan P, Liang J, Peng J. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics; 2018. pp. 2357–2368. https://aclanthology.org/D18-1258. Accessed 1 Jan 2023.
29. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25:30–6.
30. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In: Proceedings of the 4th Machine Learning for Healthcare Conference, vol. 106. PMLR; 2019. pp. 359–380.
31. Jia Y, McDermid J, Lawton T, Habli I. The Role of Explainability in Assuring Safety of Machine Learning in Healthcare. IEEE Trans Emerg Top Comput. 2022;10(4):1746–60. https://doi.org/10.1109/TETC.2022.3171314.
32. Kobo O, Brown SA, Nafee T, Mohamed MO, Sharma K, Istanbuly S, et al. Impact of malignancy on In-hospital mortality, stratified by the cause of admission: An analysis of 67 million patients from the National Inpatient Sample. Int J Clin Pract. 2021;75(11):e14758.
33. Savaliya V, Bhatnagar A, Bhavsar N, Singh M. Innovators@SMM4H'22: An Ensembles Approach for Stance and Premise Classification of COVID-19 Health Mandates Tweets. In: Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. Gyeongju: Association for Computational Linguistics; 2022. pp. 126–129. https://aclanthology.org/2022.smm4h-1.35. Accessed 1 Jan 2023.
34. Khanpour H, Caragea C. Fine-Grained Emotion Detection in Health-Related Online Posts. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics; 2018. pp. 1160–1166. https://aclanthology.org/D18-1147. Accessed 1 Jan 2023.
35. Jiang Z, Levitan SI, Zomick J, Hirschberg J. Detection of Mental Health from Reddit via Deep Contextualized Representations. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis. Online: Association for Computational Linguistics; 2020. pp. 147–156. https://aclanthology.org/2020.louhi-1.16. Accessed 1 Jan 2023.
36. Kulkarni A, Hengle A, Kulkarni P, Marathe M. Cluster Analysis of Online Mental Health Discourse using Topic-Infused Deep Contextualized Representations. In: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis. Online: Association for Computational Linguistics; 2021. pp. 83–93. https://aclanthology.org/2021.louhi-1.10. Accessed 1 Jan 2023.
37. Alimova I, Tutubalina E. Detecting Adverse Drug Reactions from Biomedical Texts with Neural Networks. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence: Association for Computational Linguistics; 2019. pp. 415–421. https://aclanthology.org/P19-2058. Accessed 1 Jan 2023.
38. Portelli B, Lenzi E, Chersoni E, Serra G, Santus E. BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics; 2021. pp. 1740–1747. https://aclanthology.org/2021.eacl-main.149. Accessed 1 Jan 2023.
39. Mesbah S, Yang J, Sips RJ, Valle Torre M, Lofi C, Bozzon A, et al. Training Data Augmentation for Detecting Adverse Drug Reactions in User-Generated Content. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics; 2019. pp. 2349–2359. https://aclanthology.org/D19-1239. Accessed 1 Jan 2023.
40. Miftahutdinov Z, Tutubalina E. Deep Neural Models for Medical Concept Normalization in User-Generated Texts. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence: Association for Computational Linguistics; 2019. pp. 393–399. https://aclanthology.org/P19-2055. Accessed 1 Jan 2023.
41. Santurkar S, Tsipras D, Madry A. Breeds: Benchmarks for subpopulation shift. 2020. arXiv preprint arXiv:2008.04859.
42. Shim H, Lowet D, Luca S, Vanrumste B. An exploratory data analysis: the performance differences of a medical code prediction system on different demographic groups. In: Proceedings of the 4th Clinical Natural Language Processing Workshop. Seattle: Association for Computational Linguistics; 2022. pp. 93–102. https://aclanthology.org/2022.clinicalnlp-1.10/.
43. Holderness E, Cawkwell P, Bolton K, Pustejovsky J, Hall MH. Distinguishing clinical sentiment: The importance of domain adaptation in psychiatric patient health records. 2019. arXiv preprint arXiv:1904.03225.
44. Wang Z, Qu Y, Chen L, Shen J, Zhang W, Zhang S, et al. Label-aware double transfer learning for cross-specialty medical named entity recognition. 2018. arXiv preprint arXiv:1804.09021.
45. Liu M, Han J, Zhang H, Song Y. Domain adaptation for disease phrase matching with adversarial networks. In: Proceedings of the BioNLP 2018 workshop. Melbourne: Association for Computational Linguistics; 2018. pp. 137–141. https://aclanthology.org/W18-2315.
46. Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Trans Neural Netw Learn Syst. 2020;32(11):4793–813.
47. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017. arXiv preprint arXiv:1702.08608.
48. Beckh K, Müller S, Jakobs M, Toborek V, Tan H, Fischer R, et al. Harnessing Prior Knowledge for Explainable Machine Learning: An Overview. In: First IEEE Conference on Secure and Trustworthy Machine Learning. 2023. pp. 450–463. In Press
49. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: An overview for clinical practitioners

– Saliency-based XAI approaches. Eur J Radiol. 2023;162:110787. https://doi.org/10.1016/j.ejrad.2023.110787.
50. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. Eur J Radiol. 2023;162:110786. https://doi.org/10.1016/j.ejrad.2023.110786.
51. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Dec Making. 2020;20:1–9.
52. Shen H, Huang TH. Explaining the Road Not Taken. 2021. arXiv preprint arXiv:2103.14973.
53. Jacovi A, Goldberg Y. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 4198–4205. https://aclanthology.org/2020.acl-main.386.
54. Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Comput Surv. 2023. https://doi.org/10.1145/3583558.
55. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–50.
56. Van Aken B, Herrmann S, Löser A. What Do You See in this Patient? Behavioral Testing of Clinical NLP Models. In: Proceedings of the 4th Clinical Natural Language Processing Workshop. Seattle: Association for Computational Linguistics; 2022. pp. 63–73. https://aclanthology.org/2022.clinicalnlp-1.7.
57. Shwartz V, Choi Y. Do Neural Language Models Overcome Reporting Bias? In: Scott D, Bel N, Zong C, editors. Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics; 2020. pp. 6863–6870. https://aclanthology.org/2020.coling-main.605. Accessed 1 Mar 2023.
58. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. 2018;20(5):e185. https://doi.org/10.2196/jmir.9134.
59. Khadzhynov D, Schmidt D, Hardt J, Rauch G, Gocke P, Eckardt KU, et al. The Incidence of Acute Kidney Injury and Associated Hospital Mortality. Deutsches Ärzteblatt Int. 2019. https://doi.org/10.3238/arztebl.2019.0397.
60. Liu F, Ge S, Wu X. Competence-based Multimodal Curriculum Learning for Medical Report Generation. 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. pp. 3001–3012. https://doi.org/10.18653/v1/2021.acl-long.234.
61. Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org; 2019. http://www.fairmlbook.org. Accessed 1 Mar 2023.
62. Chen J, Berlot-Attwell I, Wang X, Hossain S, Rudzicz F. Exploring Text Specific and Blackbox Fairness Algorithms in Multimodal Clinical NLP. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. pp. 301–312. https://aclanthology.org/2020.clinicalnlp-1.33.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.