Natural language processing in healthcare: applications and value

Buchem, M.M. van

Part 3

General Discussion and Summary

Chapter 9

General Discussion

The general aim of this thesis was to explore the application and potential value of natural language processing (NLP) in healthcare. More specifically, we studied the development and validation of NLP models in different settings, showing promising results but also highlighting challenges that need to be overcome for responsible implementation in clinical practice. Furthermore, we assessed the value of two NLP models in a pilot study. In this chapter, the main findings will be discussed and recommendations for further research will be provided.

9.1 Promising applications

NLP provides many opportunities for application in healthcare. We examine which NLP tasks hold most promise for healthcare settings. We then shift our focus to the different data sources within the healthcare environment, analyzing how NLP can be applied to these datasets to improve care delivery. An overview of promising applications is presented in Table 9.1.

9.1.1 Opportunities of various NLP tasks

Classification

Classification is the most common task in clinical natural language processing[1], which is also reflected in the studies presented in this thesis. The models we developed vary in performance, which may be due to several causes, of which the amount of training data is an important one**.** To be able to generalize to new data, a model needs to have seen a representative number of examples per class during training. The classification model for distinguishing negative from non-negative patient experiences (Chapter 4) had a lower performance than the model for distinguishing neutral from positive, which might be explained by the low number of examples for the negative experiences. Thus, we should be critical about the feasibility of training a classification model from scratch in settings with a low number of examples per label. However, clear guidelines on how much data is needed is lacking. Recent advances in NLP, such as the introduction of pretrained language models, have made it possible to train models with less training data[2–4]. The latest transformer models, such as GPT3 and later versions and Llama2, even allow for zero-shot and few-shot learning, where the model is provided zero training data or just a few examples[5,6].

Topic modeling

In Chapters 4 and 6 we had large, unlabeled datasets of patient experiences and social media data, that were subject to change over time. Topic modeling proved useful in both these settings: in the patient experiences setting (Chapter 4 and 7), it led to the identification of three new action points by the care team; in the statin setting (Chapter 6), the primary topics aligned with previous findings, such as low adherence due to adverse effects or preference for lifestyle alternatives. However, some topics uncovered novel points of discourse, such as the role of statins in improving COVID-19 outcomes, which could be interesting for future research. Within healthcare, multiple studies have used topic modeling to uncover patient perspectives from social media data with similar findings[7–11].

Other settings where topic modeling has proven useful are, for example, automatic phenotyping of patients using EHR data[12,13]. However, topic modeling is mostly useful for data exploration and as part of an analysis pipeline. Without any postprocessing, the topic descriptions are difficult to interpret and thus not suitable for settings where a clear label is needed. Often, the resulting topics are used as input to a classification model, using topic modeling for dimensionality reduction[12,13]. The recent introduction of large language models (LLMs) might offer a solution to the lack of interpretability by being able to describe the contents of the different topics[14]. This might improve the applicability of topic modeling on text data in a healthcare context. Furthermore, large language models might be able to take over the full topic modeling pipeline, as these models have unprecedented summarization capabilities. To date, only one study performed this comparison. This study shows that LLMs can be a robust substitute for current topic modeling techniques, although extensive prompt engineering is necessary to extract useful topics[15].

Summarization

The task of summarizing texts has seen a huge jump in performance with the introduction of LLMs. At the time of writing the scoping review (Chapter 2), most studies used extractive summarization, entity extraction, and classification to filter and sort the information relevant for a summary, although this did not lead to natural, narrative summaries. The introduction of LLMs has made

all these previous techniques obsolete for summarization purposes. Our pilot study with Autoscriber (Chapter 8) shows the high quality of these summaries: medical students are positively surprised by the quality of the summaries and a majority wants to use this technique in their work. In the past few months, several similar studies have been published, showing promising performance of LLMs in creating discharge summaries[16] and clinical information letters for patients[17,18], and summarizing radiology reports[19–21], patient-doctor dialogues[20,22–24], and the many handovers between shifts or departments happening within the hospital every day[25]. Many of these opportunities address the current administrative burden in healthcare, while potentially improving the quality of the collected data because of increased consistency, structure, and objectiveness. These opportunities are not without challenges and risks, which will be discussed in Section 9.3.1. Furthermore, the major challenge related to summarization is the difficulty in creating a reference standard. This topic will be discussed in Section 9.2.1.

9.1.2 Opportunities of various data sources

In this thesis, we focused on three data types: clinical data, patient-generated data, and social media data. Within the literature, clinical data such as EHR data is the most common data source used within the clinical NLP field[26]. As most information about the patient is captured within free text EHR data, a large focus within the clinical NLP field has been to use this data to develop patient risk classification and prediction models[27]. In our acute care utilization setting (Chapter 3) we show that models trained on only free-text data perform similarly to models trained only on structured data, without the need for extensive preprocessing. This finding is in line with other studies, underlining the value of free text data relative to structured data[28,29]. A future research question would be if the use of free text EHR data allows for easier transfer of models between organizations. There are, however, many challenges related to the use of EHR data, which we will discuss in Section 9.2.1. A new type of clinical data is the recorded and transcribed clinical conversation. With the introduction of large language models, it has become possible to leverage the value of this data type. Apart from the previously discussed summarization capabilities of these models, the application of LLMs on clinical conversational data leads to a

new level of granularity in clinical data. Having complete transcripts of clinical conversations might lead to more, in-depth and structured data about disease trajectories and patient presentations of various diseases, potentially paving the way for improved clinical decision support. No literature is available on this topic yet.

Another popular data source due to its accessibility is social media data. We investigated different ways in which social media data might improve clinical practice in our statin setting (Chapter 6) and patient messages setting (Chapter 5). In the former, social media was used to gain insights into public views on statins. The results of this study can be used to inform clinicians which topics to discuss when talking to patients about statins. In the latter, we used continued pretraining on Reddit data, aiming to improve the performance of a pretrained language model on identifying depression concerns from patient messages. In this case, this process did not improve the overall performance of the model. With respect to direct value for healthcare, social media allows for extracting patient perspectives, coping strategies, and even adverse drug events[30–33]. Furthermore, social media data is being used to pretrain (large) language models such as Llama 2[5].

A relatively underused data type is patient-generated data, such as patient experience data and patient messages. The potential benefits of this data type are underlined in our patient experience and patient messages settings (Chapters 3, 4, and 7). We see that capturing free-text patient experiences leads to new insights, which can be turned into concrete points of improvement, thus having the potential to directly improve patients' experiences. In Chapter 5, we see that patient messages show signs of distress that can be used to identify depression concerns. These settings show the value of this data type, along with the value of using NLP to take away the burden of analyzing all the incoming data from healthcare professionals. Especially patient messages are underused, while these could be very valuable in creating triggers for mental health issues or other acute care needs. Recent studies mostly highlight the burden, potential value, and characteristics of this data type, but examples in clinical practice are lacking[34–37]. Recent studies use LLMs to create patient-focused chatbots or

create draft replies to patient messages, with mixed results [38,39]. Although it might have the potential to make healthcare organizations more accessible to patients and free up time from healthcare professionals, this has not been shown in large-scale evaluations.

Table 9.1: an overview of the most promising applications of NLP in healthcare.

Data source & task	Application
Clinical data & summarization	Use NLP models with summarization capabilities to automatically generate clinical notes, discharge letters, and summaries of patient files.
Clinical data & classification	Use BERT and LLMs to train classifiers in settings with a small amount of training data. Train classification models on free-text data to simplify the transfer of models between organizations.
Patient-generated data & classification	Apply classification models to patient messages to create triggers for mental health or acute care needs.
Patient-generated data & summarization	Use NLP models with summarization capabilities to provide relevant information to patients.
Social media data & topic modeling	Use topic modeling techniques in combination with LLMs to extract general perspectives about health-related topics from social media data.

9.2 Challenges during development

In the course of developing NLP models for the settings outlined in Chapters 3 to 6, we faced multiple challenges that hindered their adoption in clinical settings. This section details these challenges, including issues related to data, bias, explainability, and deployment. See Figure 1 for an overview.

9.2.1 Data

As with other types of machine learning, NLP models are only as good as the data they are trained on. Especially in healthcare, access to large amounts of high-quality data is a huge challenge. This challenge encompasses the quality of the data itself, its availability and accessibility, and the choice of a trustworthy reference standard.

Data quality

Most clinical NLP models are trained on EHR data, which is not primarily collected for research or development purposes but to create a complete record of clinical information and proceedings[40]. The lack of standardized methods to create free-text clinical notes in combination with the wide variety in the human use of language leads to large heterogeneity of free-text clinical data between clinicians, departments, and organizations[40]. A recent study reported differences in medical coding practices between two organizations due to a different billing system, affecting clinicians' incentive to code[41]. Furthermore, diagnostic coding systems such as ICD-10 are primarily used for billing purposes and thus may not always correspond to the actual diagnosis[42]. Other studies report high rates of missing problems in the problem list, especially if physicians are overstretched[43,44]. In our acute care utilization case (Chapter 3), an extra preprocessing step was added to reduce the large number of redundant notes that were found in patients' records[45,46]. However, NLP might provide a solution to this challenge. Many studies use NLP to improve the quality and structure of EHR data, such as extracting the presence of symptoms or the medication dosage and linking terms to ontologies (e.g. UMLS[47], ICD-10[48], or SNOMED-CT[49])[50].

Data availability

An essential aspect of developing NLP tools for healthcare is having data available for training purposes. Within the healthcare context, data availability can be challenging due to the challenges around data sharing. This holds for structured data, but even more so for text data as it is not always possible to fully anonymize the data. As seen in Chapters 2 through 8, this means most studies include data from a single institution. This trend is seen across the whole clinical NLP field, with most studies using proprietary data[1,50]. In the past few years, many solutions have been proposed to tackle this challenge. Federated learning is a technique where the machine learning model is trained on data from multiple organizations, but instead of the data leaving the organizations to be collected in a central location, the model 'travels' from organization to organization[51]. Another promising solution is the development of synthetic data, where a model is trained to create new data that is very similar to the training data. Using this

technique, a hospital can create a new dataset that is shareable with other organizations, without sharing actual patient data[52,53].

Reference standard

Reference standards are essential for supervised NLP methods requiring labeled data, but also for evaluating the output of unsupervised methods. While some settings will have very clear reference standards (such as in-hospital mortality), finding a trustworthy reference standard is a challenge in many settings. Apart from reference standard issues related to data quality and data availability as discussed in the previous sections, other challenges arise when a reference standard is not present. In our patient messages setting (Chapter 5) we manually labeled a large dataset of patient messages as concerning or not concerning. We defined 'concerning' with a group of clinicians and created an extensive annotation guideline. Even so, the inter-annotator agreement was 0.38, which is considered moderate. Such a score is not uncommon in the healthcare domain. A study describing the development of a large corpus of annotated clinical conversations reports similar scores, especially for entities such as symptoms and conditions[54]. The topic of inter-annotator agreement or inter-rater reliability has been described repeatedly in medical literature[55,56]. However, it takes on a different meaning when these labels are not just used for quality control, but for training AI models that might take over tasks from trained clinicians. Although in some cases a single truth can be unveiled, in many cases trained clinicians will have irreconcilable disagreements about ground truth labels. A growing body of research recognizes the importance of including these differing opinions in the development process to improve the generalizability of the model and, most importantly, decrease bias[57,58]. Furthermore, a recent paper argues that machine learning in healthcare should strive to move past the quality and accuracy of current clinical practice[59]. They recommend using objective ground truth metrics such as mortality or patient centered metrics such as perceived pain and, where possible, taking ambitious steps to elevate existing standards.

9.2.2 Bias

With the current large-scale development of NLP models in healthcare, we must pay attention to who is benefiting from these models and who is potentially

harmed. Especially when using patient-generated data, we must be conscious of the differences in access to certain communication methods and patients' ability to express themselves. In the evaluation of our patient experience model (Chapter 7), we found that patients who filled out the questionnaire had a higher level of education. Furthermore, of the patients that completed the questionnaire, patients who were male and who had a low level of education more often responded with only one word. In our patient messages setting (Chapter 5), our cohort consisted of mostly White and Asian, privately insured patients. In this last example, we risk perpetuating existing biases in the (lack of) recognition of depression in different populations. Both examples show the need to critically reflect on who has access to the tools we develop[60–62].

A more recent problem in relation to bias is the data that (large) language models are trained on. Our own RedditBERT model (Chapter 5) was trained on Reddit data, which is not representative for patients in general. Autoscriber (Chapter 8) uses GPT3.5, which has a bias towards dominant values from the United States because of the available data for training[63]. A recent paper reviewed the current clinical language models and found that almost all of them were trained on the MIMIC-III dataset[64]. This dataset contains data from the intensive care unit of the Beth Israel Deaconess Medical Center, an academic medical center in Boston, which will not be representative for large parts of healthcare[59]. There are many ways to mitigate bias, including guidance ethics, increasing diversity of AI developers, and increasing diversity within the data. Guidance ethics is a method developed in The Netherlands that uses a multistakeholder perspective to uncover the potential effects of the model, aiming to strengthen the positive effects while mitigating the negative effects[65]. An essential part of guidance ethics is including a representative and diverse group of people to include different perspectives, which is also important for the AI community as a whole. We discuss this topic in our commentary 'Picture a data scientist', providing several recommendations[66]. Lastly, several initiatives are working on creating datasets that include a diverse representation of the population. The All of Us research program is a great example, with more than 80% of its participants belonging to groups that have historically been underrepresented in biomedical research[67].

9.2.3 Clinically relevant metrics

Many papers presenting NLP models do not publish performance metrics that are clinically relevant. Within the machine learning community, it is standard practice to report metrics such as the precision and recall based on the performance of the model on an unseen test set. Obviously, these statistical performance metrics do not consider how the model will be used in clinical practice and if the output of the model is clinically useful and which actions can subsequently be taken[42,59,68]. Furthermore, these metrics do not provide a comprehensive view of whether an NLP model will deliver value in healthcare settings, making it challenging to discern which models are promising for further development and which ones may not be useful. For every setting, researchers should critically assess which performance metrics are relevant and in line with clinical goals and workflows. Good examples of clinically relevant metrics include decision curve analysis[69] and the number needed to benefit[70]. A recent paper proposed an extensive evaluation framework for healthcare chatbots, including an evaluation of the interface and interaction with the user, and highlighting the importance of manual evaluations by end users[71]. Ultimately, value must be evaluated in a clinical trial, which will be discussed in Section 9.3.3. However, the metrics suggested here can provide valuable insights during development.

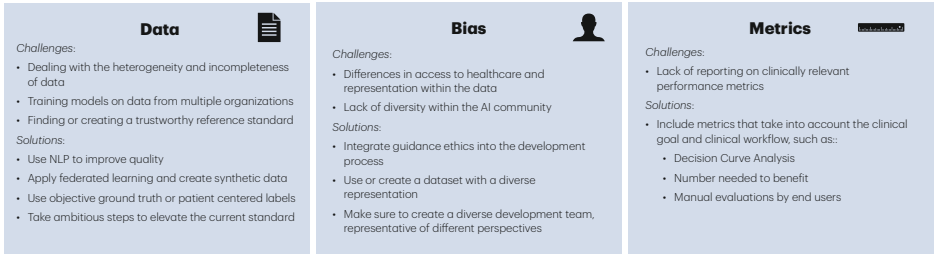

Figure 1: an overview of the challenges during development of NLP models for healthcare.

9.3 Value for clinical practice

Although the promise of applying natural language processing to healthcare is undeniable and vast amounts of NLP tools are developed, scientific reports on the value of these tools when deployed in clinical practice are lacking. Our

scoping review (Chapter 2) highlights this in the realm of digital scribes, and similar patterns are observed in other areas[72]. Among the models discussed in Chapters 2 through 6, two have been developed into applications, either by integration with existing systems or as standalone tools. This section discusses the definition of value for clinical practice, how to create value with NLP applications, and how to evaluate this in clinical practice (see Figure 2).

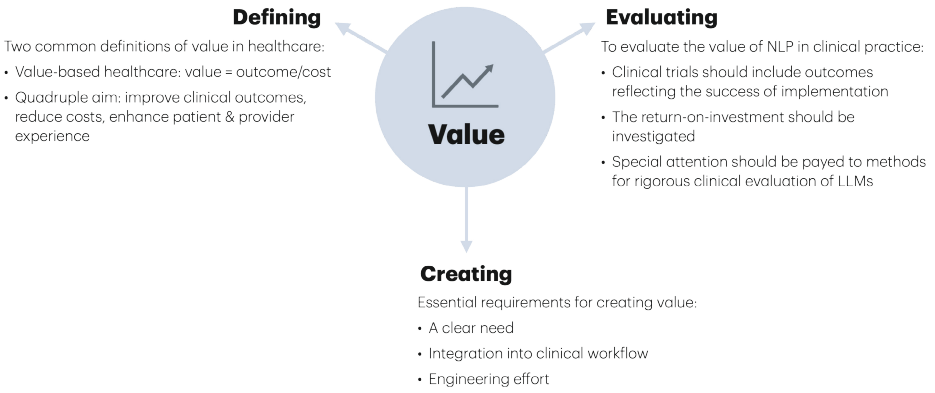

Figure 2: an overview of defining, creating, and evaluating value.

9.3.1 Defining Value in Healthcare

A first step in investigating the value of NLP for healthcare is defining what value for healthcare encompasses. The most widespread approach to value for healthcare has been the value-based healthcare (VBHC) approach introduced by Porter and Teisberg[73]. They argue that value within healthcare should be defined around the patient and can be calculated as follows:

$$Value = \frac{Outcome}{Costs}$$

These outcomes go beyond traditional metrics such as medical procedures or tests and encompass a broader spectrum of factors such as patient-reported outcomes (PROs), functional status, quality of life, and long-term health outcomes. Thus, VBHC focuses on weighing the impact of healthcare interventions

on patients' health status and overall well-being against their costs. Another common approach towards assessing value within healthcare is the Quadruple Aim[74,75]. This approach describes value in similar terms as VBHC, including patient outcomes, costs, and patient experience, but also adds an extra dimension: clinician experience (see Figure 3).

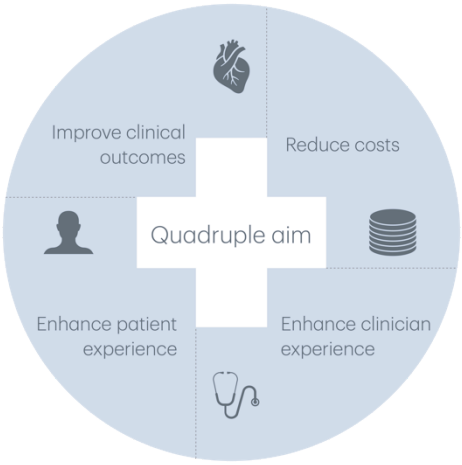

Figure 3: the four dimensions of the Quadruple Aim.

In terms of these definitions of value for healthcare, the promise of NLP lies in several dimensions. The first and most straightforward dimension is the clinician experience. With an increasing rate of clinician burnout and decreased job satisfaction, there is an urgent need for improvement[76–80]. With the promise of taking over tasks such as writing discharge letters, generating clinical notes, and replying to patient queries, NLP might have a serious impact [81]. Furthermore, several studies have shown the ability of NLP to perform diagnostic tasks[82,83]. These developments could lead to a decrease in resources needed for diagnostic work-up. It is still speculative at this point in time whether costs and patient outcomes improve, since large, high-quality clinical trials are lacking.

9.3.2 Creating Value with NLP Applications

Within this thesis, we did not reach the clinical trial stage with any of the NLP models. However, we performed two pilot studies, and together with findings

from the literature they highlight essential requirements for creating value with NLP applications: a clear need, integration into the clinical workflow, and engineering effort.

A clear need

Both the AI-PREM (Chapter 7) and Autoscriber (Chapter 8) originated from clear needs from clinical practice. As pointed out in recent research, it is crucial to distinguish between 'machine learning on healthcare data' and 'machine learning for healthcare problems'[59]. While the former advances the field theoretically, it is the latter that tends to create actual value in healthcare. Developers and researchers need to step away from primarily developing models for settings where large datasets and labels are readily available and engage with healthcare professionals to find the right problems[42]. This is a fundamental step towards realizing clinical value.

Integration into the clinical workflow

A recent commentary states that the key to achieving meaningful impact with AI is focusing on behavioral change and thus on changing routines and care processes[84]. For the AI-PREM and Autoscriber, significant effort was invested in determining how they would fit into the existing clinical workflows. This process involved extensive discussions with a multidisciplinary team over many iterations. In the past few years, a few useful guidelines have been published, describing the process from idea to implementation in healthcare[85,86]. These guidelines specifically pay attention to integrating the model into the clinical workflow.

Engineering effort

Another essential requirement for workflow integration is the engineering effort needed to develop an application or integrate into an existing application. For the AI-PREM, we chose to integrate it into an already existing dashboard, while Autoscriber is currently a standalone web application that will soon be able to integrate with the EHR. These engineering efforts demand considerable effort and expertise and are currently often neglected in development projects[59]. To achieve widespread clinical value, it is essential to simplify these engineer-

ing steps. In radiology, for example, several companies have begun offering 'AI platforms' that provide plug-and-play functionality for various AI applications, facilitating large-scale integration into hospital systems[87,88].

9.3.3 Evaluating Value

Since high-quality, large-scale clinical trials with NLP models are lacking, it is important to reflect on how to best evaluate the value of these models in clinical practice[81]. As discussed in our scoping review (Chapter 2), many studies lack error analyses, and clinical validation or utility assessments are almost non-existent. These assessments are paramount in gaining clinicians' trust in these models. A few recent perspectives provide guidance on how to evaluate AI in clinical practice. First, clinical trials should include outcomes that reflect the success of implementation[89]. These include the compatibility with the workflow, the adoption rate, and the cost of implementation, which are crucial for understanding the success or failure of an AI tool.

Furthermore, with the current shortages in healthcare, an essential aspect of AI tools to consider during evaluation is the return-on-investment. A recent paper developed a return-on-investment calculator to inform decision making for an AI-powered radiology diagnostic imaging platform[90]. The calculator compares the current workflow to the updated workflow after deployment of the AI tool and provides insights into aspects such as time savings, effects on clinical outcomes, healthcare services provided, and the total cost. Of course, assumptions need to be made for aspects where data is unavailable. However, insights on the return-on-investment of AI tools will become increasingly important with a growing supply of tools.

Lastly, the introduction of LLMs poses considerable challenges to the evaluation of these tools due to their generative nature. Challenges include the lack of clearly defined evaluation metrics, variation in the design of human evaluations, and incorporating the human-LLM interaction into the evaluation[68]. In response to this, we recently published guidance, describing three tiers of clinical LLM validation (see Box 1)[91]. There are two recent examples of rigorous evaluations of LLMs in healthcare. The first study evaluated the value of

a tool that automatically generates notes from clinical conversations[23]. The tool was implemented in clinical practice and available for 10.000 physicians and staff. The authors evaluated the output of the tool automatically and manually, investigated the effect of the tool on several EHR logging metrics (e.g. time spent in the EHR after working hours), and collected patient and physician experiences. The second study evaluated several adapted LLMs on four clinical summarization tasks[92]. Apart from performing similar extensive automatic and manual evaluations of the summaries, they also investigated the potential medical threat that errors could pose. These two studies provide important examples on how to rigorously evaluate LLMs in clinical settings.

Box 1. Three tiers of medical Large Language Model validation[91]

> **1. General validation**
> General validation assesses general LLM quality independent of the performed task. Important outcomes at this stage may be the LLM's robustness to different formulations of the same prompt and the readability of the LLM output.
> **2. Task specific validation**
> Task specific validation assesses the LLM performance on task specific outcomes. For example, for summarization it may be the consistency with source material and coverage of important clinical concepts.
> **3. Clinical validation**
> Clinical validation assesses the LLM performance on specific healthcare outcomes. The validation goals at this tier will depend on the clinical objectives and intended use, such as improved health outcomes, higher patient satisfaction or reduction in administration time.

9.4 Future outlook

9.4.1 Large language models

The field of natural language processing has changed inconceivably over the past few years. Recent studies describe the promising performance of LLMs

in extracting information, drafting responses to patients, summarizing patient records, and even for prediction tasks[68,93]. A notable breakthrough was reported in a recent study by Google, where an AI system was capable of conducting a comprehensive diagnostic (chat) dialogue with patients, achieving diagnostic accuracy that surpassed that of trained physicians[83]. While LLMs have demonstrated significant potential, the translation of these capabilities into practical healthcare benefits remains largely unproven. Moreover, LLMs introduce several new challenges that need attention. These challenges include their generative nature and the hallucinations this may lead to, the energy consumption necessary for training and inference, and concerns about data privacy[64,94]. The latter might be even more pressing than before, because of the accessibility of tools such as ChatGPT. Educating healthcare professionals on the basics of AI and how they might use these tools is crucial to ensure appropriate and safe use.

Given these challenges and the current uncertainties inherent in new technologies, we should keep a critical stance towards the use of LLMs while also recognizing the potential opportunities they may offer. There are some interesting examples of tools that use LLMs as part of their NLP pipeline, using their language generation capabilities in combination with information retrieval models or curated datasets to control the possible outputs[95,96]. As stated in Section 9.3.4, such tools should be rigorously evaluated before use in clinical practice. It is clear, however, that the introduction of large language models has completely changed the field of NLP and will greatly impact how we practice healthcare in the coming years.

9.4.2 Role of companies

As discussed in our scoping review (Chapter 2), the current role of commercial companies within the field is ambivalent. On one hand, companies are often not transparent about the development and validation of their models. On the other hand, they play an essential role in turning promising technologies into software applications that can be used in clinical practice. With legislation such as the EU's Medical Device Regulation and the AI-act, this process is lengthy and expensive. Furthermore, during this process input from many different experts is needed.

To make sure promising NLP models end up as reliable software products, good collaboration between researchers and companies is needed. Creating a pipeline where healthcare organizations work on proof-of-concepts, which, if successful, are further developed into software products by companies could be beneficial to all parties involved. Validation and impact assessment by independent researchers should be a mandatory part of this pipeline, recognizing the dynamic nature of the field with continuous improvements.

Furthermore, electronic health records (EHR) vendors should play an important role in these developments. Since the launch of ChatGPT, EHR vendor Epic has greatly invested in integrating this technique into their EHR[97]. However, the largest Dutch EHR vendor, has yet to facilitate easy integration of AI. Governments and healthcare organizations should actively think about what this landscape should look like and the accompanying demands this should place on EHR vendors.

9.5 Recommendations

For hospitals	For developers
Develop a strategy for AI in general. This should include objectives for the use of AI, roles and responsibilities for development and deployment, and policy on how and when to engage with companies.	Be steered by the needs from healthcare, instead of the availability of data.
Invest in robust infrastructure to facilitate the deployment of NLP tools.	Include clinical expertise from the start to get in-depth knowledge about the meaning of the data and the clinical workflow.
Include all relevant expertise for every development and implementation project. If you do not have this expertise internally, make sure to involve independent external experts.	Critically reflect on the necessity to use complex NLP models. Sometimes less is more.
Educate your healthcare professionals on the basics of LLMs, how to use them responsibly, and the associated risks.	Rigorously evaluate your NLP models using established reporting standards and evaluation frameworks.

9.6 Conclusion

The field of natural language processing has seen inconceivable progress over the past five years, and with it the possibilities for application in healthcare. Although we identify many promising applications in this thesis, challenges related to data quality and availability, bias, and a lack of insights in clinically relevant metrics remain. These challenges hinder the further development or implementation of many NLP models in healthcare. To turn NLP models into valuable additions for clinical practice, we should pay more attention to working on the right problems, reporting on clinically relevant metrics, lowering the engineering effort needed to integrate a model into the clinical workflow, and performing thorough clinical impact evaluations. If these challenges are addressed, NLP may significantly improve clinician experience, patient experiences and outcomes, reduce costs, and keep healthcare accessible and affordable.

References

1. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, Turner K. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. Comput Biol Med 2023;155:106649. PMID:36805219
2. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arXiv 2019; PMID:31501885
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. Arxiv 2017;
4. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Arxiv 2018;
5. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv 2023; doi: 10.48550/arxiv.2307.09288
6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. arXiv 2020; doi: 10.48550/arxiv.2005.14165
7. Anjum A, Zhao X, Bahja M, Lycett M. Identifying patient experience from online resources via sentiment analysis and topic modelling. Proc 3rd Ieee Acm Int Conf Big Data Comput Appl Technologies 2016;94–99. doi: 10.1145/3006299.3006335
8. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel Approach to Cluster Patient-Generated Data Into Actionable Topics: Case Study of a Web-Based Breast Cancer Forum. Jmir Medical Informatics 2018;6(4):e45. PMID:30497991
9. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, Asch DA, Ungar LH, Merchant RM. Yelp Reviews Of Hospital Care Can Supplement And Inform Traditional Surveys Of The Patient Experience Of Care. Health Affair 2017;35(4):697–705. PMID:27044971
10. Cammel SA, Vos MSD, Soest D van, Hettne KM, Boer F, Steyerberg EW, Boosman H. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. Bmc Med Inform Decis 2020;20(1):97. PMID:32460734
11. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. Bmj Heal Care Informatics 2021;28(1):e100262. PMID:33653690

12. Jiang H, Huang X, Zhang J, Song Z, Toral XS, Xu Y, Liu A, Guo L, Powell G, Verma A, Buckeridge D, Marelli A, Li Y. Supervised multi-specialist topic model with applications on large-scale electronic health record data. Proc 12th ACM Conf Bioinform, Comput Biol, Heal Inform 2021;1–26. doi: 10.1145/3459930.3469543
13. Ahuja Y, Zou Y, Verma A, Buckeridge D, Li Y. MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. J Biomed Inform 2022;134:104190. PMID:36058522
14. Grootendorst M. LLM & Generative AI. Available from: https://maartengr.github.io/BERTopic/getting_started/representation/llm.html [accessed May 3, 2024]
15. Mu Y, Dong C, Bontcheva K, Song X. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. arXiv 2024; doi: 10.48550/arxiv.2403.16248
16. Williams CYK, Bains J, Tang T, Patel K, Lucas AN, Chen F, Miao BY, Butte AJ, Kornblith AE. Evaluating Large Language Models for Drafting Emergency Department Discharge Summaries. medRxiv 2024;2024.04.03.24305088. PMID:38633805
17. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Heal 2023;5(4):e179–e181. PMID:36894409
18. Roberts RHR, Ali SR, Dobbs TD, Whitaker IS. Can Large Language Models Generate Outpatient Clinic Letters at First Consultation That Incorporate Complication Profiles From UK and USA Aesthetic Plastic Surgery Associations? Aesthetic Surg J Open Forum 2023;6:ojad109. PMID:38192329
19. Ma C, Wu Z, Wang J, Xu S, Wei Y, Liu Z, Jiang X, Guo L, Cai X, Zhang S, Zhang T, Zhu D, Shen D, Liu T, Li X. ImpressionGPT: An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT. arXiv 2023; doi: 10.48550/arxiv.2304.08448
20. Veen DV, Uden CV, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134–1142. PMID:38413730
21. López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, Luna A. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: A feasibility study. Int J Méd Inform 2024;187:105443. PMID:38615509
22. Yim W, Fu Y, Abacha AB, Snider N, Lin T, Yetisgen M. Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. Sci Data 2023;10(1):586. PMID:37673893
23. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, Liu V, Lee K. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM Catal 2024;5(3). doi: 10.1056/cat.23.0404
24. Lyu M, Peng C, Li X, Balian P, Bian J, Wu Y. Automatic Summarization of Doctor-Patient Encounter Dialogues Using Large Language Model through Prompt Tuning. arXiv 2024; doi: 10.48550/arxiv.2403.13089
25. Sezgin E, Sirrianni J, Kranz K. Development and Evaluation of a Digital Scribe: Conversation Summarization Pipeline for Emergency Department Counseling Sessions towards Reducing Documentation Burden. medRxiv 2023;2023.12.06.23299573. PMID:38106162

26. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, Soni S, Wang Q, Wei Q, Xiang Y, Zhao B, Xu H. Deep learning in clinical natural language processing: a methodical review. J Am Méd Inform Assoc 2020;27(3):457–470. PMID:31794016

27. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, Rosand B, Li Y, Zhang M, Chang D, Taylor RA, Krumholz HM, Radev D. Neural Natural Language Processing for unstructured data in electronic health records: A review. Comput Sci Rev 2022;46:100511. doi: 10.1016/j.cosrev.2022.100511

28. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Méd Inform Decis Mak 2020;20(1):280. PMID:33121479

29. Mugisha C, Paik I. Pneumonia Outcome Prediction Using Structured And Unstructured Data From EHR. 2020 IEEE Int Conf Bioinform Biomed (BIBM) 2020;00:2640–2646. doi: 10.1109/bibm49941.2020.9312987

30. Hollander D den, Dirkson AR, Verberne S, Kraaij W, Oortmerssen G van, Gelderblom H, Oosten A, Reyners AKL, Steeghs N, Graaf WTA van der, Desar IME, Husson O. Symptoms reported by gastrointestinal stromal tumour (GIST) patients on imatinib treatment: combining questionnaire and forum data. Support Care Cancer 2022;30(6):5137–5146. PMID:35233640

31. Dirkson A, Verberne S, Kraaij W, Oortmerssen G van, Gelderblom H. Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers. Sci Rep 2022;12(1):10317. PMID:35725736

32. Dirkson A, Verberne S, Oortmerssen G van, Gelderblom H, Kraaij W. How do others cope? Extracting coping strategies for adverse drug events from social media. J Biomed Inform 2023;139:104228. PMID:36309197

33. Jeyaraman M, Ramasubramanian S, Kumar S, Jeyaraman N, Selvaraj P, Nallakumarasamy A, Bondili SK, Yadav S. Multifaceted Role of Social Media in Healthcare: Opportunities, Challenges, and the Need for Quality Control. Cureus 2023;15(5):e39111. PMID:37332420

34. Aakre CA. Applying Natural Language Processing Neural Network Architectures to Augment Appointment Request Review of Self-Referred Patients to an Academic Medical Center. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci 2022;2022:85–91. PMID:35854757

35. Davoudi A, Lee NS, Luong T, Delaney T, Asch EL, Chaiyachati KH, Mowery DL. Identifying Medication-related Intents from a Bidirectional Text Messaging Platform for Hypertension Management: An Unsupervised Learning Approach. medRxiv 2021;2021.12.23.21268061. doi: 10.1101/2021.12.23.21268061

36. Huang M, Fan J, Prigge J, Shah ND, Costello BA, Yao L. Characterizing Patient-Clinician Communication in Secure Medical Messages: Retrospective Study. J Méd Internet Res 2022;24(1):e17273. PMID:35014964

37. Steitz BD, Sulieman L, Warner JL, Fabbri D, Brown JT, Davis AL, Unertl KM. Classification and analysis of asynchronous communication content between care team members involved in breast cancer treatment. JAMIA Open 2021;4(3):ooab049. PMID:34396056

38. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, Chipidza FE, Leeman J, Aerts HJWL, Miller T, Savova GK, Gallifant J, Celi LA, Mak RH, Lustberg M, Afshar M, Bitterman DS. The effect of using a large language model to respond to patient messages. Lancet Digit Heal 2024; PMID:38664108

39. Tai-Seale M, Baxter SL, Vaida F, Walker A, Sitapati AM, Osborne C, Diaz J, Desai N, Webb S, Polston G, Helsten T, Gross E, Thackaberry J, Mandvi A, Lillie D, Li S, Gin G, Achar S, Hofflich H, Sharp C, Millen M, Longhurst CA. AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. JAMA Netw Open 2024;7(4):e246565. PMID:38619840

40. Edmondson ME, Reimer AP. Challenges Frequently Encountered in the Secondary Use of Electronic Medical Record Data for Research. CIN: Comput, Inform, Nurs 2020;38(7):338–348. PMID:32149742

41. Yim W-W, Wheeler AJ, Curtin C, Wagner TH, Hernandez-Boussard T. Secondary use of electronic medical records for clinical research: challenges and opportunities. Converg Sci Phys Oncol 2018;4(1):014001. PMID:29732166

42. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN, Thadaney-Israni S, Goldenberg A. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337–1340. PMID:31427808

43. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: An audit of problem list completeness during the COVID-19 pandemic. Int J Méd Inform 2021;150:104452. PMID:33864979

44. Wang EC-H, Wright A. Characterizing outpatient problem list completeness and duplications in the electronic health record. J Am Méd Inform Assoc 2020;27(8):1190–1197. PMID:32620950

45. Searle T, Ibrahim Z, Teo J, Dobson R. Estimating redundancy in clinical text. J Biomed Inform 2021;124:103938. PMID:34695581

46. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and Sources of Duplicate Information in the Electronic Medical Record. JAMA Netw Open 2022;5(9):e2233348. PMID:36156143

47. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(suppl_1):D267–D270. PMID:14681409

48. Organization WH. ICD-10 : international statistical classification of diseases and related health problems : tenth revision. 2004. Available from: https://iris.who.int/handle/10665/42980 ISBN:9241546549

49. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Heal Technol Inform 2006;121:279–90. PMID:17095826

50. Oliveira JM de, Costa CA da, Antunes RS. Data structuring of electronic health records: a systematic review. Heal Technol 2021;11(6):1219–1235. doi: 10.1007/s12553-021-00607-w

51. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020;2(6):305–311. doi: 10.1038/s42256-020-0186-1

52. Guan J, Li R, Yu S, Zhang X. A Method for Generating Synthetic Electronic Medical Record Text. IEEEACM Trans Comput Biol Bioinform 2019;18(1):173–182. PMID:31647443

53. Zhou N, Wu Q, Wu Z, Marino S, Dinov ID. DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes. J Méd Syst 2022;46(12):96. PMID:36380246

54. Shafran I, Du N, Tran L, Perry A, Keyes L, Knichel M, Domin A, Huang L, Chen Y, Li G, Wang M, Shafey LE, Soltau H, Paul JS. The Medical Scribe: Corpus Development and Model Performance Analyses. Arxiv 2020.

55. Oommen C, Howlett-Prieto Q, Carrithers MD, Hier DB. Inter-Rater Agreement for the Annotation of Neurologic Concepts in Electronic Health Records. medRxiv 2022;2022.11.16.22282384. doi: 10.1101/2022.11.16.22282384

56. Bajpai S, Bajpai RC, Chaturvedi HK. Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods. Journal of the Indian Academy of Applied Psychology 2015;41(No.3 (Special Issue)):20–27.

57. Gordon ML, Lam MS, Park JS, Patel K, Hancock J, Hashimoto T, Bernstein MS. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. CHI Conf Hum Factors Comput Syst 2022;1–19. doi: 10.1145/3491102.3502004

58. Raghu M, Blumer K, Sayres R, Obermeyer Z, Kleinberg R, Mullainathan S, Kleinberg J. Direct Uncertainty Prediction for Medical Second Opinions. arXiv 2018; doi: 10.48550/arxiv.1807.01771

59. Balagopalan A, Baldini I, Celi LA, Gichoya J, McCoy LG, Naumann T, Shalit U, Schaar M van der, Wagstaff KL. Machine learning for healthcare that matters: Reorienting from technical novelty to equitable impact. PLOS Digit Heal 2024;3(4):e0000474. PMID:38620047

60. Timmons AC, Duong JB, Fiallo NS, Lee T, Vo HPQ, Ahle MW, Comer JS, Brewer LC, Frazier SL, Chaspari T. A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health. Perspect Psychol Sci 2023;18(5):1062–1096. PMID:36490369

61. Raza S, Garg M, Reji DJ, Bashir SR, Ding C. Nbias: A natural language processing framework for BIAS identification in text. Expert Syst Appl 2024;237:121542. doi: 10.1016/j.eswa.2023.121542

62. Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, Eydelman MB, Foundation W DC Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging, Maisel WH. Considerations for addressing bias in artificial intelligence for health equity. npj Digit Med 2023;6(1):170. PMID:37700029

63. Johnson RL, Pistilli G, Menédez-González N, Duran LDD, Panai E, Kalpokiene J, Bertulfo DJ. The Ghost in the Machine has an American accent: value conflict in GPT-3. arXiv 2022; doi: 10.48550/arxiv.2203.07785

64. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. npj Digit Med 2023;6(1):135. PMID:37516790

65. Verbeek P-P, Tijink D. Guidance ethics approach: An ethical dialogue about technology with perspective on actions. ECP | Platform voor de InformatieSamenleving; 2020. Available from: https://ris.utwente.nl/ws/portalfiles/portal/247401391/060_002_Boek_Guidance_ethics_approach_Digital_EN.pdf [accessed May 2, 2024]

66. Hond AAH de, Buchem MM van, Hernandez-Boussard T. Picture a data scientist: a call to action for increasing diversity, equity, and inclusion in the age of AI. J Am Méd Inform Assoc 2022;29(12):2178–2181. PMID:36048021

67. Investigators A of URP. The "All of Us" Research Program. N Engl J Med 2019;381(7):668–676. PMID:31412182

68. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA 2023;330(9):866–869. PMID:37548965

69. Vickers AJ, Calster B van, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3(1):18. PMID:31592444

70. Liu VX, Bates DW, Wiens J, Shah NH. The number needed to benefit: estimating the value of predictive analytics in healthcare. J Am Méd Inform Assoc 2019;26(12):1655–1659. PMID:31192367

71. Abbasian M, Khatibi E, Azimi I, Oniani D, Abad ZSH, Thieme A, Sriram R, Yang Z, Wang Y, Lin B, Gevaert O, Li L-J, Jain R, Rahmani AM. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. npj Digit Med 2024;7(1):82. PMID:38553625

72. Sande D van de, Genderen ME van, Huiskens J, Gommers D, Bommel J van. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensiv Care Med 2021;47(7):750–760. PMID:34089064

73. Porter ME, Teisberg EO. Redefining Health Care: Creating Value-based Competition on Results. Harvard Business Review Press; 2006. Available from: https://books.google.nl/books?id=cse2LOAndNIC ISBN:9781422133361

74. Bodenheimer T, Sinsky C. From Triple to Quadruple Aim: Care of the Patient Requires Care of the Provider. Ann Fam Med 2014;12(6):573–576. PMID:25384822

75. Sikka R, Morath JM, Leape L. The Quadruple Aim: care, health, cost and meaning in work. BMJ Qual Saf 2015;24(10):608. PMID:26038586

76. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV, Carlasare LE, Dyrbye LN. Changes in Burnout and Satisfaction With Work-Life Integration in Physicians and the General US Working Population Between 2011 and 2017. Mayo Clin Proc 2019;94(9):1681–1694. PMID:30803733

77. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W-J, Sinsky CA, Gilchrist VJ. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Medicine 2017;15(5):419–426. PMID:28893811

78. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, Westbrook J, Tutty M, Blike G. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med 2016;165(11):753. PMID:27595430

79. Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, Wang W, Luft HS. Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine. Health Affair 2017;36(4):655–662. PMID:28373331

80. Rao SK, Kimball AB, Lehrhoff SR, Hidrue MK, Colton DG, Ferris TG, Torchiana DF. The Impact of Administrative Burden on Academic Physicians. Acad Med 2017;92(2):237–243. PMID:28121687

81. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930–1940. PMID:37460753

82. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, Deng A, Baston A, Ross J, Idowu E, Teo JT, Dobson RJ. Foresight -- Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs. arXiv 2022; doi: 10.48550/arxiv.2212.08072

83. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S, Singhal K, Cheng Y, Hou L, Webson A, Kulkarni K, Mahdavi SS, Semturs C, Gottweis J, Barral J, Chou K, Corrado GS, Matias Y, Karthikesalingam A, Natarajan V. Towards Conversational Diagnostic AI. arXiv 2024; doi: 10.48550/arxiv.2401.05654

84. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care. JAMA 2019;321(23):2281–2282. PMID:31107500

85. Faneyte S. Stappenplan Healthy AI (HAI). Maasstad Ziekenhuis; Available from: https://www.nvki.nl/community/threads/stappenplan-healthy-ai.471/ [accessed May 12, 2024]

86. Handelingsruimte A. Tool Handelingsruimte Waardevolle AI voor Gezondheid en Zorg. Dutch Ministry for Health, Welfare and Sport; Available from: https://nlaic.com/wp-content/uploads/2022/06/04a.-Hulpmiddel-Handelingsruimte-Waardevolle-AI-voor-gezondheid-en-zorg.pdf [accessed May 12, 2024]

87. Sectra Amplifier Services: AI adoption made easy. Available from: https://medical.sectra.com/product/sectra-amplifier-services/ [accessed May 12, 2024]

88. Blackford Analysis. Available from: https://blackfordanalysis.com [accessed May 12, 2024]

89. Sande D van de, Chung EFF, Oosterhoff J, Bommel J van, Gommers D, Genderen ME van. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. npj Digit Med 2024;7(1):58. PMID:38448743

90. Bharadwaj P, Nicola L, Breau-Brunel M, Sensini F, Tanova-Yotova N, Atanasov P, Lobig F, Blankenburg M. Unlocking the Value: Quantifying the ROI of Hospital AI. J Am Coll Radiol 2024; PMID:38499053

91. Hond A de, Leeuwenberg T, Bartels R, Buchem M van, Kant I, Moons KG, Smeden M van. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Heal 2024;6(7):e441–e443. PMID:38906607

92. Veen DV, Uden CV, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134–1142. PMID:38413730

93. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023;388(25):2399–2400. PMID:37342941

94. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. Proc 57th Annu Meet Assoc Comput Linguistics 2019;3645–3650. doi: 10.18653/v1/p19-1355

95. Consensus. AI Search Engine for Research. Available from: https://consensus.app/home/about-us/ [accessed May 13, 2024]

96. Glass.Health. AI-Powered Clinical Decision Support. Available from: https://glass.health [accessed May 13, 2024]

97. Center MN. Microsoft and Epic expand strategic collaboration with integration of Azure OpenAI Service. 2023. Available from: https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/ [accessed May 2, 2024]