Study ID	Title	Objective	Methods	Results	Themes
Abu-Ashour 2024	Using Artificial Intelligence to Label Free-Text Operative and Ultrasound Reports for Grading Pediatric Appendicitis.	Comparing data extraction quality between ChatGPT-4 and human data collectors for structuring free-text EMR data in pediatric appendicitis.	Humans extracted preoperative and operative data from 2,100 children. ChatGPT-4 prompted to structure 103 anonymized reports using set variables and coding options. A pediatric surgeon adjudicated the outputs.	Manual extraction accounted for 98% of errors, with 28.2% misclassification vs 2.9% by ChatGPT-4. AI avoided misclassification in 59.2% of records, ~40 times faster.	Item Identification
Acharya 2024	Clinical risk prediction using language models: benefits and considerations.	Study use of language models for EHR-based risk prediction in data-limited settings by integrating supplementary domain knowledge.	"LLaMA2-EHR" and "Sent-e-Med" to utilize textual EHR descriptions for predicting future diagnoses. Compared them with established methods across various data types and sizes.	LMs delivered significant improvements in ROC and PR metrics, enabled few-shot learning, and managed unseen medical concepts, though prompt sensitivity remained a notable factor.	Clinical Predcition
Afshar 2024	On the role of the UMLS in supporting diagnosis generation proposed by Large Language Models.	Three objectives: Probe the medical knowledge of LLMs. Examine the benefit of grounding them with UMLS, and assess correlations between UMLS-based metrics and human judgments.	Diagnoses from health questions and care notes were evaluated using ConsumerQA and Problem Summarization datasets. Comparing prompts with or without UMLS paths and analyzing multiple metrics.	GPT-3.5 surpassed Llama2 in one-hop UMLS path completion (F1=10. 9%). Grounding with UMLS raised SapBERT scores by 4%, and standard metrics (ROUGE and SapBERT) had reduced correlation with expert evaluations.	Clinical Decision Making
Alkhalaf 2024	Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records.	Test the efficacy of zero-shot prompt engineering, alone and with retrieval augmented generation (RAG), for summarizing structured/unstructured EHR data and extracting malnutrition information in 40 Australian RACFs.	The Llama 2 13B model was applied with zero-shot prompting to unstructured and structured EHRs, first independently, then with RAG, to generate structured summaries of nutritional status and identify malnutrition risk factors.	Summaries achieved 93.25% accuracy, increasing to 99.25% with RAG, and risk factor extraction reached 90% accuracy with no further gains from RAG.	Data Extraction
Beaulieu-Jones 2023	Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study.	Assess machine learning models, including large language models (LLMs), to see if they can predict seizure recurrence after a first-time seizure-like event in children.	A retrospective cohort study (2010–2020) compared structured-data models (logistic regression, XGBoost) with LLM-based approaches on two datasets. 1. Boston Children's Hospital EMRs (14 021 patients) and 2. IBM MarketScan claims (15 062 patients).	LLMs with domain-specific and location-specific pre-training achieved the best performance (F1=0.826, AUC=0.897), outperforming the base LLM (F1=0.739, AUC=0.846) and structured-data methods (best F1=0. 679, AUC=0.725).	Clinical Predcition
Bi 2024	Adapting Large Language Models for Automated Summarisation of Electronic Medical Records in Clinical Coding.	Investigate LLMs, particularly those pre-trained on biomedical data, to see if they can effectively summarize clinical narratives by adapting to the clinical domain.	A dataset of EHR text paired with "Brief Hospital Course" from MIMIC-III was used to fine-tune two open-source LLMs. One was pre-trained on biomedical data and the other was trained on a general domain. Then evaluated against their base models.	The model pre-trained on biomedical data achieved better summarization performance after clinical adaptation, suggesting that domain-specific pre-training may be advantageous for assisting clinical coders.	Summarization
Bridges 2024	Computerized diagnostic decision support systems - a comparative performance study of Isabel Pro vs. ChatGPT4.	Validate the diagnostic accuracy of ChatGPT4 by comparing its diagnosis lists to Isabel Pro.	201 cases were tested using identical inputs. The mean reciprocal rank (MRR) was used to measure the correctness and rank of diagnoses, and a Wilcoxon Signed Rank Sum test assessed system equivalence.	Isabel Pro accuratly identified 175 diagnoses (87.1%) vs 165 (82.1%) by ChatGPT4, with MRR scores of 0.428 and 0.389, respectively.	Clinical Decision Making
Chiang 2024	A large language model-based generative natural language processing framework fine-tuned on clinical notes accurately extracts headache frequency from electronic health records.	Study the performance of a GPT-2 generative model to extracting headache frequency from free-text EHR notes	This retrospective cross-sectional study used 1,915 neurology consultation notes from three Mayo Clinic sites (2012–2022). Compared four NLP approaches (ClinicalBERT and three GPT-2-based models) for accuracy and R^2 in predicting headache frequency.	The GPT-2 generative model achieved the highest accuracy (0.92) and R^2 (0.89). It outperformed ClinicalBERT and showed accurate performance for diverse clinical notes.	Data Extraction
Cho 2024	Extracting lung cancer staging descriptors from pathology reports: A generative language model approach.	Fine-tuned generative language models (GLMs) evaluated for extracting key lung cancer staging information from surgical pathology reports.	42 descriptors required for tumor-node (TN) classification was validated by two clinical experts, and the resulting gold standard was used to create prompt-response pairs to train and test six GLMs on lung cancer pathology reports.	The Deductive Mistral-7B model demonstrated the best performance: 92.24% exact match ratio in information extraction and an accuracy of 0.9876 for concurrent T and N classification.	Data Extraction
Chung 2024	Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication.	Examine GPT-4 Turbo's predictive performance on 8 perioperative tasks	This prognostic study used task-specific datasets from two years of retrospective EHR data, formatting case notes into prompts for GPT-4 Turbo. Classification tasks evaluated with F1 scores, and duration predictions were assessed with mean absolute error.	GPT-4 Turbo achieved its highest F1 scores of 0.81 for ICU admission and 0.86 for hospital mortality. Duration predictions had errors of 49 minutes for PACU phase 1 and 4.5 days for hospital length of stay.	Clinical Predcition
Doshi 2024	Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis.	Compare four LLMs, ChatGPT-3.5, ChatGPT-4, Bard Gemini, and Bing in simplifying radiology report impressions for patient comprehension.	750 anonymized radiology report impressions from the MIMIC-IV database were used. 3 prompt variants were tested, and simplification was evaluated via four established readability indexes. Wilcoxon signed-rank tests assessed differences in reading grade levels.	All LLMs lowered the reading grade level of radiology report impressions ($P < .001$), with further reductions when given patient context or requested to simplify at the seventh-grade level. Performance varied based on the prompt used.	Patient Communication
Elmarakeby 2023	Empirical evaluation of language modeling to ascertain cancer outcomes from clinical text reports.	Evaluate 9 NLP models under different training strategies for extracting cancer response and progression outcomes from imaging reports in non-small cell lung cancer.	A dataset of 14,218 imaging reports from 1,112 patients was used. A BERT-based model (DFCI-ImagingBERT) further pre-trained on 662,579 unlabeled reports from 27,483 patients.	The DFCI-ImagingBERT generally produced the best results, but simpler models and convolutional neural network models were only slightly behind in performance.	Data Extraction
Fraser 2023	Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study.	Comparing the performance of ChatGPT3.5 and 4.0, WebMD, and Ada symptom checkers in diagnosing and triaging urgent or emergent clinical conditions, using emergency department final diagnoses and physician reviews as reference points.	Deidentified self-report data from 40 ED patients who had used the Ada symptom checker were entered into ChatGPT3.5 and 4.0 and WebMD by a blinded research assistant. The proportion of top 1 and top 3 diagnostic matches was calculated, and triage recommendations were compared with physician ratings.	Top 1 diagnostic match rates for Ada, ChatGPT 3.5, ChatGPT 4.0, and WebMD were 30%, 40%, 33%, and 40%, respectively (47% for physicians). ChatGPT 3.5 demonstrated higher unsafe triage rates (41%) than Ada (14%), whereas ChatGPT 4.0 had lower diagnostic accuracy, better triage agreement (76% agreement with physicians). Ada and WebMD performed better overall than ChatGPT.	Clinical Decision Making
Fu 2024	CACER: Clinical concept Annotations for Cancer Events and Relations.	Study relationship between cancer drugs and associated symptom burdens by extracting from oncology clinical notes.	CACER was created with over 48,000 medical problems and drug events plus 10,000 related event relations. Transformer-based models (BERT, Flan-T5, Llama3, and GPT-4) were evaluated using fine-tuning and in-context learning.	Fine-tuned BERT and Llama3 achieved performance comparable to inter-annotator agreement (around 88 F1) for event extraction. This outperformed GPT-4's in-context learning approach. BERT performed similar to Llama3.	Data Extraction
Gabriel 2024	On the development and validation of large language model-based classifiers for identifying social determinants of health.	Evaluate LLM-based classifiers for identifying social determinants of health from clinical text.	Synthetic training data were generated by combining authentic clinical notes with outputs from generative pre-trained transformers, which were then used to train BERT and RoBERTa models. Validated on both MIMIC-III and institutional EHR data.	On the institutional dataset, the combined synthetic and authentic approach achieved area under the ROC curve values of 0.78 for homelessness, 0.72 for food insecurity, and 0.83 for domestic violence.	Item Identification

Gandomi 2024	ARDSFlag: an NLP/machine learning algorithm to visualize and detect high-probability ARDS admissions independent of provider recognition and billing codes.	Develop an algorithm to automate ARDS detection and create a visualization tool to assist clinicians in assessing ARDS criteria.	ARDSFlag combines machine learning and NLP to evaluate structured and unstructured EHR data (MIMIC-III), using separate text classifiers validated on 300 cases blindly labeled by clinicians.	ARDSFlag achieved 89.0% accuracy (specificity=91.7% recall=80.3% precision=75.0%) in detecting ARDS.	Clinical Predcition						
Garcia 2024	Artificial Intelligence-Generated Draft Replies to Patient Inbox Messages.	Evaluate LLM for generating draft responses to patient portal messages	Prospective quality improvement study at a single academic medical center (5 weeks). 162 enrolled clinicians received AI-generated draft replies. Outcome measures included utilization rate, time metrics (reply action, write, and read times), and clinician experience.	Clinicians used AI-generated drafts for 20% of patient message replies on average. No significant changes in time metrics were observed. Physician taskload and work exhaustion scores significantly decreased from presurvey to postsurvey	Patient Communication						
Ge 2024	A Comparison of a Large Language Model vs Manual Chart Review for the Extraction of Data Elements From the Electronic Health Record.	Comparing GPT-4 LLM (Versa GPT-4) with manual chart review for extracting eight hepatocellular carcinoma (HCC) imaging data elements.	1,101 pre- and post–locoregional treatment imaging reports from 753 patients in the UCSF FrAILT study were manually annotated for distinct HCC-related features. These were processed in 56 runs of a few-shot prompt to Versa GPT-4.	Versa GPT-4 achieved an overall accuracy of 0.934 relative to manual chart review, ranging from 0.886 (sum of tumor diameters) to 0.989 (extrahepatic metastases). Accuracy was generally higher for classification tasks.	Data Extraction						
Glicksberg 2024	Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room.	Evaluate GPT-4's performance in predicting emergency department admissions from real-world data compared to traditional ML models. Also explore if few-shot examples, retrieval-augmented generation (RAG), and/or ML numeric probabilities can enhance GPT-4's predictive capabilities	Using electronic health records from 7 NYC hospitals. Bio-Clinical-BERT (unstructured data) and XGBoost (structured data) were trained, combined into an ensemble, and then compared with GPT-4 in multiple scenarios: Zero-shot, Few-shot (with and without RAG), and with and without ML numeric probabilities	The ensemble ML model achieved an AUC of 0.88, AUPRC of 0.72, and accuracy of 82.9%; GPT-4's performance (0.79 AUC, 0.48 AUPRC, 77.5% accuracy) improved to near-ML levels (up to 0.87 AUC, 0.71 AUPRC, 83.1% accuracy) when supplemented with RAG and/or ML probabilities.	Clinical Predcition						
Gu 2024	Automatic quantitative stroke severity assessment based on Chinese clinical named entity recognition with domain-adaptive pre-trained large language model.	Develop a framework for quantitatively assessing stroke severity by directly scoring the NIHSS from Chinese EHRs.	A Chinese stroke clinical record dataset was created. A domain-adaptive pre-trained LLM called CliRoberta was used. Extracted entities were mapped to NIHSS items, allowing for an automated scoring pipeline	The domain-adaptive LLM–based CNER model achieved an F1 of 0.990 on extracting entities. The fully automated NIHSS scoring showed strong agreement (0.823) and high consistency (0.986) with the ground truth.	Clinical Prediction, Clinical Decision Making						
Hao 2024	A GPT-based EHR modeling system for unsupervised novel disease detection.	Develop an AI-based anomaly detection model to identify atypical hospital cases and detect emerging outbreaks by analyzing EHR time sequences	A GPT-based system was trained using Empirical Risk Minimization on 31 clinical variables from 120,714 hospitalizations to produce an Out-Of-Distribution (OOD) anomaly score. Comparison between this and traditional linear models looking at patient-level anomaly detection and mortality prediction	In an unsupervised setting, the GPT model detected SARS-CoV-2 infections with AUC 92.2%, surpassing linear models in both anomaly detection (AUC=78.3%) and mortality prediction (AUC=94.7%)	Clinical Predcition						
Hu 2024	Zero-shot information extraction from radiological reports using ChatGPT.	Explores if ChatGPT can accurately extract specific information from radiological reports in a zero-shot manner	A prompt template was developed for CT reports. ChatGPTs responses were transformed into structured results using a post-processing module. Additional medical knowledge was added to some prompt templates to explore if this could reduce errors.	ChatGPT achieved competitive performance on tasks like tumor location and diameter extraction compared with a baseline system. Incorporating prior medical knowledge significantly improved performance on certain tasks but did not enhance accuracy for more complex extractions such as tumor density or lymph node status.	Data Extraction						
Iscoe 2024	Identifying signs and symptoms of urinary tract infection from emergency department clinical notes using large language models.	Introduce NLP fundamentals to the emergency medicine research community and develop two task-specific LLMs to identify UTI symptoms from unstructured ED notes	A random subset of 1,250 ED clinician notes (annotated for 17 UTI symptoms) was divided into a training set of 1,000 notes and a 250 note test set to compare a convolutional neural network–based model (SpaCy) and a transformer-based model (Clinical Longformer)	The transformer model achieved an overall F1 measure of 0.88 compared to 0.84 for SpaCy when identifying note-level UTI symptoms. For detecting the presence or absence of any symptom, Clinical Longformer attained an F1 of 0.98 versus 0.96 with SpaCy.	Item Identification						
Jiang 2023	Health system-scale language models are all-purpose prediction engines.	Demonstrate that LLMs trained on unstructured clinical notes (NYUTron) can serve as clinical predictive engines	A clinical language model was trained on a large amount of electronic health record notes and fine-tuned for five prediction tasks (30-day all-cause readmission, in-hospital mortality, comorbidity index, length of stay, and insurance denial). Model performance was evaluated using the area under the curve (AUC), and the system was used in a prospective, single-arm trial	NYUTron achieved AUCs ranging from 78.7% to 94.9%, providing a 5.36–14.7% improvement over traditional models. Clinical text pretraining and fine-tuning to specific sites improved performance and generalizability, demonstrating the potential of clinical language models to assist at the point of care	Clinical Predcition						
Kaufmann 2024	Validation of a Zero-shot Learning Natural Language Processing Tool to Facilitate Data Abstraction for Urologic Research.	Develop and validate a zero-shot learning NLP tool based on GPT-3.5 to facilitate data abstraction from unstructured text in urologic research	The zero-shot learning NLP tool based on GPT-3.5 was compared to three physicians for speed and accuracy on 199 deidentified radical prostatectomy pathology reports, processing both vectorized and scanned documents	Physicians needed a median 93 seconds/report versus 12-15 seconds for the software, which achieved 94.2% accuracy on vectorized reports and 88.7% on scanned reports, demonstrating noninferiority to at least two out of three human abstractors	Summarization						
Lee 2024	Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage.	Develop and validate LLMs for automatic EHR recording in the ED by identifying 12 medical symptoms and 2 patient histories from simulated clinician–patient conversations	4 transformer-based pre-trained models were fine-tuned for classification. Explanations were generated using eXplainable AI (XAI) and the SHAP method. A Turing test compared the XAI outputs with explanations from medical workers. An emergency medicine specialist evaluated both	The KLUE-RoBERTa-based model achieved the highest performance (F1=0.965, AUROC=0.893), and its SHAP-based explanations showed a high similarity (Jaccard=0.722) to those of medical workers. The Turing test revealed only a 6% gap between XAI and human explanations.	Item Identification						
Lin 2024	An Advanced Machine Learning Model for a Web-Based Artificial Intelligence-Based Clinical Decision Support System Application: Model Development and Validation Study.	Develop and validate an advanced ML model for breast cancer recurrence prediction within a web-based AI-CDSS	The Tri-Service General Hospital breast cancer registry (N=3577, 2004–2016) provided training (2004–2012) and validation (2013–2016) cohorts. ChatGPT facilitated data cleaning and feature engineering was included. Multiple algorithms (e.g., light gradient boosting, gradient boosting, extreme gradient boosting) were evaluated. Synthetic minority oversampling technique addressed class imbalance.	Light gradient boosting had the highest performance (AUC=0.80), closely followed by other gradient boosting methods. A functional web-based AI-CDSS demonstrated potential for personalized treatment planning and patient engagement.	Clinical Decision Making						
Liu 2024	Leveraging large language models for generating responses to patient messages-a subjective analysis.	Develop and assess the performance of fine-tuned LLMs for generating responses to patient messages on an electronic health record patient portal	A pre-trained LLaMA-65B model was fine-tuned on over 499,000 patient–provider message pairs (CLAIR-Short). It was then combined with updated, more educational responses from an open-source dataset (CLAIR-Long). 4 primary care physicians evaluated the generated responses from CLAIR-Short, CLAIR-Long, and ChatGPT for empathy, responsiveness, accuracy, and usefulness	CLAIR-Short produced concise, provider-like replies. CLAIR-Long provided more patient education, earning ratings similar to ChatGPT in empathy, responsiveness, and accuracy but neutral in usefulness	Patient Communication						
Maciejewski 2024	AssistMED project: Transforming cardiology cohort characterisation from electronic health records through natural language processing - Algorithm design, preliminary results, and field prospects.	Develop and validate AssistMED, an NLP data extraction tool for EHRs.	AssistMED was designed to retrieve patient clinical conditions, medications, and diagnostic results. Its performance was compared with manual data retrieval on a sample of 400 cardiology inpatients.	AssistMED accurately analyzed 56 clinical conditions, medications from 16 drug groups with dosages, and 15 echocardiographic parameters with no significant differences compared to manual retrieval. Minor discrepancies were mainly caused by a combination of random algorithm errors, human annotation mistakes, and the tool's limited context awareness.	Data Extraction						
Morales-Sanchez 2024	Early diagnosis of HIV cases by means of text mining and machine learning models on clinical notes.	Evaluate the effectiveness of LLMs in classifying patients as suspected or not of having HIV using unstructured clinical notes from EHRs.	Compiled a dataset of clinical notes from patients classified as either HIV suspects or HIV non-suspects, and compared classical machine learning algorithms with two biomedical Spanish LLMs.	Both LLMs outperformed classical ML algorithms, achieving F-scores of 94.7 in unbalanced settings and RoBERTa reaching an F-score of 95.7 in balanced settings - ultimately indicating that biomedical LLMs are able to effectively and efficiently reduce missed opportunities for HIV diagnosis.	Item Identification						

Musacchio 2024	A transparent machine learning algorithm uncovers HbA1c patterns associated with therapeutic inertia in patients with type 2 diabetes and failure of metformin monotherapy.	Identify and categorize determinants that influence the intensification of therapy in T2DM patients with suboptimal blood glucose control (despite metformin monotherapy).	Utilized the Logic Learning Machine to analyze 1.5 million patients' EHRs from 271 Italian diabetes clinics between the years of 2005 and 2019. Focused on patients on metformin monotherapy with two consecutive mean HbA1c levels greater than 7 %, categorizing them into "inertia-NO" and "inertia-YES".	The LLM achieved a ROC-AUC of 0.81, accuracy of 0.71, precision of 0.80, recall of 0.71, and an F1 score of 0.75. It identified two distinct subtypes of therapeutic inertia. The first one being a gradual steady increase in HbA1c and the other being a moderate, fluctuating rise.	Item Identification						
Pathak 2023	Extracting Thyroid Nodules Characteristics from Ultrasound Reports Using Transformer-based Natural Language Processing Methods.	Develop and evaluate transformer-based NLP models (such as GatorTron), to extract thyroid nodule characteristics from ultrasound reports.	After a multidisciplinary team created annotation guidelines and a labeled corpus, five transformer models (BERT, RoBERTa, LongFormer, DeBERTa, and GatorTron) were evaluated and compared against each other on their ability to extract 16 thyroid nodule characteristics from ultrasound reports.	GatorTron achieved the highest strict and lenient F1-scores of 0.8851 and 0.9495 for extracting 16 characteristics, and an F1-score of 0.9321 for linking characteristics to nodules - outperforming all other transformer-based models.	Data Extraction						
Reynolds 2024	Comparing the quality of ChatGPT- and physician-generated responses to patients' dermatology questions in the electronic medical record.	Determine whether ChatGPT can generate high-quality responses to patient-submitted dermatology questions.	Extracted patient questions and corresponding dermatologist responses from the EMR. Inputted questions into ChatGPT-3.5 and subsequently had ten blinded reviewers (a combination of seven physicians and three nonphysicians) rate responses on overall quality, readability, accuracy, thoroughness, and empathy.	Of 31 analyzed messages, physician-generated responses were vastly preferred over ChatGPT-generated ones, and received higher ratings in both readability and level of empathy.	Patient Communication						
Rosen 2024	Evaluating the reliability of ChatGPT as a tool for imaging test referral: a comparative study with a clinical decision support system.	Determine whether ChatGPT can generate imaging referral recommendations that are as accurate as those provided by the ESR iGuide.	Conducted a comparative study with 97 emergency department abdominal cases, comparing imaging referrals from ESR iGuide and ChatGPT. Analyzed disagreements and had four specialists grade the appropriateness of ChatGPT's chest abdominal pelvis CT recommendations in 66 cases.	ChatGPT's recommendations were largely consistent with the ESR iGuide. There was no noticed difference based on age or gender. Additionlly there was high agreement with specialists. The small amount of disagreements involved vague recommendations.	Clinical Decision Making						
Roy 2024	Large-scale identification of social and behavioral determinants of health from clinical notes: comparison of Latent Semantic Indexing and Generative Pretrained Transformer (GPT) models.	Evaluate ML methods for detecting social and behavioral determinants of health from EHR data.	Applied Latent Semantic Indexing and both GPT-3.5 and GPT-4 models to over 2 million MIMIC-III clinical notes to identify 15 SBDH categories, using logistic regression with demographic and ICD-9 data for predictions.	LSI achieved a macro-F1 of 0.74, outperforming ICD-9 and GPT-3.5 but trailing GPT-4. When combined with ICD-9, LSI and GPT models had similar macro-F1 scores.	Item Identification						
Scott 2024	Assessing Artificial Intelligence-Generated Responses to Urology Patient In-Basket Messages.	Evaluate the quality and appropriateness of ChatGPT-generated responses to urology patient messages.	Generated responses for 100 categorized patient messages using ChatGPT. 5 urologists independently then rated each response on accuracy, completeness, harmfulness, helpfulness, intelligibility, and sendability.	47% of ChatGPT responses were deemed acceptable for sending to patients. Responses to easy questions were acceptable in 56% of cases versus 34% for difficult questions (P = .03), and were rated higher in accuracy, completeness, helpfulness, and intelligibility. Additionally, no differences were found based on question content.	Patient Communication						
Seng 2024	Tracking persistent postoperative opioid use: a proof-of-concept study demonstrating a use case for natural language processing.	Demonstrate that an NLP engine can accurately identify patients with persistent postoperative opioid use after major spine surgery.	Reviewed clinical documents of 965 patients who underwent major spine surgery, using an NLP engine to detect opioid use ≥3 months post-surgery. Then results were compared using manual clinician reviews.	The NLP engine accurately determined opioid use status in 92.9% of cases (correctly identifying persistent use in 95.6% and no persistent use in 86.1% of patients).	Data Extraction						
Seth 2024	Perforator Selection with Computed Tomography Angiography for Unilateral Breast Reconstruction: A Clinical Multicentre Analysis.	Compare unilateral breast reconstruction perforator selection preferences between six experienced plastic surgeons and four LLMs using CTA images.	Six plastic surgeons from various countries reviewed ten CTA images to choose the abdomen side and reconstruction type. ChatGPT-4o, ChatGPT-4, Bing CoPilot, Claude were prompted to perform the same selections. Preferences were first averaged and then compared.	Surgeons predominantly recommended DIEP (83%), with minimal use of TRAM flaps. LLMs ChatGPT-4o, ChatGPT-4, and Bing CoPilot exclusively recommended DIEP, while Claude suggested DIEP in 90% and MS TRAM in 10% of cases. Both surgeons and LLMs consistently preferred DIEP. Interestingly, LLMs occasionally deviated in their recommendations.	Clinical Decision Making						
Seto 2024	Trial of Converting Nursing Records into Simplified and Structured Information Utilizing ChatGPT-3.5.	Using ChatGPT-3.5 to convert descriptive records into concise bulleted formats.	Analyzed 96,000 anonymized nursing records and employed ChatGPT-3.5 to extract keywords and generate bulleted records based on common categories such as status on return, temperature control, blood drainage, and more.	ChatGPT-3.5 demonstrated usable functionality for extracting keywords in bulleted records, indicating potential for facilitating the transition to standardized nursing records.	Summarization						
Shapiro 2024	Application of a natural language processing artificial intelligence tool in psoriasis: A cross-sectional comparative study on identifying affected areas in patients' data.	Compare ChatGPT-4's ability to identify psoriasis-affected body areas from EMR data with that of a senior dermatologist.	Analyzed EMR data from 94 psoriasis patients at Sheba Medical Center using ChatGPT-4 to identify affected body areas, comparing the results of the LLM to the dermatologist annotations.	ChatGPT-4 correctly identified 92.8% of affected areas and achieved complete accuracy in 57.4% of instances.	Data Extraction						
Sivarajkumar 2024	Extraction of sleep information from clinical notes of Alzheimer's disease patients using natural language processing.	Automate the extraction of specific sleep-related patterns from clinical notes of Alzheimer's disease patients to investigate their role in alzheimer's incidence and progression.	Manually annotated 570 clinical notes for 12 sleep categories from the adSLEEP corpus. Developed and compared rule-based NLP, machine learning models, and LLM-based algorithms to extract relevant data.	In a dataset of 482 alzheimer's patients, the rule-based NLP algorithm achieved the highest F1 scores across all sleep-related concepts. ML models and LLAMA2 fine-tuned models performed best in specific categories. Interestingly, LLM-based approaches demonstrated promising results despite limited sleep information in the training data.	Data Extraction						
Small 2024	Large Language Model-Based Responses to Patients' In-Basket Messages.	**Assess primary care physicians' perceptions of GenAI drafts.**	**16 primary care doctors reviewed a total of 344 patient messages and rated them on the quality of their information content, communication quality, and usability.**	**GenAI responses scored higher in communication style and empathy compared to HCP responses. GenAI responses maintained similar information content and usability, but used more subjective and positive language and was more linguistically complex.**	Patient Communication						
Soroudi 2024	Comparing Provider and ChatGPT Responses to Breast Reconstruction Patient Questions in the Electronic Health Record.	**Evaluate ChatGPT's effectiveness in answering breast reconstruction queries.**	**10 deidentified breast reconstruction questions were answered by ChatGPT-3, ChatGPT-4, plastic surgeons, and advanced practice providers. Responses were rated for accuracy and empathy using Likert scales and assessed for readability with the Flesch Reading Ease score.**	**Provider responses were more readable than LLM responses. However, Chatgpt demonstrated higher empathy scores. There was no significant differences in accuracy between LLM responses and those of providers.**	Patient Communication						
Sushil 2024	A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports.	**Evaluate whether LLMs can reduce the need for extensive data annotations in extracting information in the context of breast cancer pathology reports.**	**Compared the zero-shot classification performance of GPT-4, GPT-3.5, Starling, and ClinicalCamel against supervised models (random forests, LSTM-Att, UCSF-BERT) using a manually labeled dataset of 769 breast cancer pathology reports across 12 categories.**	**GPT-4 achieved an average macro F1-score of 0.86, significantly outperforming or matching the best supervised model (LSTM-Att, F1=0.75) especially on imbalanced tasks. All other LLMs testes, demonstrated poor performance.**	Item Identification						
Tan 2024	Artificial Intelligence in Hand Surgery - How Generative AI is Transforming the Hand Surgery Landscape.	**Assess both the evolution and impact of LLMs in healthcare.**	**Reviewed historical AI developments and analyzed current LLM implementations in healthcare.**	**LLMs enhance patient care and reduce physician workload by automating administrative tasks and aiding in diagnosis and surgical planning. However, many challenges including bias, hallucination, and EMR integration difficulties remain.**	Other						

vanNuland 2024	Poor performance of ChatGPT in clinical rule-guided dose interventions in hospitalized patients with renal dysfunction.	**Assess ChatGPT's effectiveness in providing medication dose adjustments in the context of patients with renal impairment.**	**Analyzed 172 CDSS alerts from Tergooi Medical Center's EHR over two weeks, comparing ChatGPT's medication recommendations with an expert panel's responses.**	**Without patient variables, ChatGPT correctly matched expert recommendations in 46.6% of alerts and with patient variables, correct matches were 32.7%.**	Clinical Decision Making							
VanVeen 2024	Adapted large language models can outperform medical experts in clinical text summarization.	Evaluate the effectiveness of adapted LLMs in various clinical summarization tasks.	Assessed LLM performance using quantitative NLP metrics and a clinical reader study with 10 physicians on four tasks:radiology reports, patient questions, progress notes, and doctor-patient dialogue	The best-adapted LLMs were rated as equivalent or superior to those from medical experts in 81% of cases, while safety analysis identified challenges such as errors and fabricated information.	Summarization							
Williams 2024	Evaluating the use of large language models to provide clinical recommendations in the Emergency Department.	Determine whether LLMs can accurately provide clinical recommendations for admission status, radiological investigations, and antibiotic prescriptions from data found in emergency department EMRs.	Evaluated 10,000 randomly selected emergency department visits by comparing zero-shot, GPT-3.5-turbo, and GPT-4-turbo-generated recommendations across four prompting strategies against resident physician recommendations.	Both GPT-4-turbo and GPT-3.5-turbo underperformed compared to resident physicians, with accuracy scores 8% and 24% lower, respectively.	Clinical Decision Making							
Woo 2024	Evaluation of GPT-4 ability to identify and generate patient instructions for actionable incidental radiology findings.	Assess GPT-4's ability to: 1) identify actionable incidental findings from Emergency Department radiology reports and 2) evaluate the quality of AI-generated patient summaries.	Manually categorized findings in 50 validation and 430 test radiology reports, fine-tuned GPT-4 prompts, and measured accuracy and summary appropriateness using a Likert scale.	GPT-4 achieved 99.3% recall and 73.6% precision for actionable findings. Interestingly, the majority of AI-generated summaries required no or minor revisions.	Item Identification							
Wu 2024	A hybrid framework with large language models for rare disease phenotyping.	Combine dictionary-based NLP tools with LLMs to enhance the identification of rare diseases from unstructured clinical reports.	Developed a rare disease vocabulary using ORDO and UMLS. SemEHR was used to extract rare disease mentions. Tested different strategies, including providing few examples or none at all, to improve accuracy.	The hybrid framework outperformed traditional NLP systems and standalone LLMs. LLaMA3 and Phi3-mini achieving the highest F1 scores. Few-shot prompting with 1-3 examples provided the best performance, while knowledge-augmented generation offered limited improvements. Additionally, the approach successfully identified numerous potential rare disease cases not documented in structured diagnostic records, demonstrating its ability to uncover previously unrecognized patients.	Item Identification							
Xie 2024	Disparities in seizure outcomes revealed by large language models.	Evaluate an epilepsy-specific LLM for intrinsic bias across demographic groups and assess seizure outcome disparities using LLM-extracted data.	The LLM was tested for differential performance based on race, ethnicity, sex, income, and insurance using manually annotated clinical notes. Seizure freedom classifications were analyzed for outcome disparities across these demographic groups in 84,675 clinic visits.	The LLM demonstrated consistent accuracy and class balance across all demographic groups.	Other							
Yan 2024	Prompt engineering on leveraging large language models in generating response to InBasket messages.	This study aimed to determine whether LLMs can generate high-quality draft responses to patient medical advice requests that satisfy both patients and clinicians.	Investigators conducted three rounds of human-involved iterative prompt engineering with GPT-4, refining its output based on clinician and patient feedback; the final prompt was then validated on independent data and implemented in a real EHR environment with 69 primary care clinicians.	Physician acceptance of LLM-generated drafts increased from 62% to 84%, while 74% of test drafts were rated "helpfu,." 76% of patients could not distinguish human from LLM-generated responses, and 72% of clinicians felt it reduced cognitive load.	Patient Communication							
Zaretsky 2024	Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format.	To determine whether a LLM can transform discharge summaries into more readable, patient-friendly language.	Fifty adult inpatient discharge summaries were processed to produce simplified versions. The outputs were then compared with the original notes using readability (Flesch-Kincaid) and understandability (PEMAT) metrics, as well as physician assessments for accuracy and completeness.	The LLM-generated summaries showed significantly lower reading levels (mean 6.2 vs 11.0) and higher understandability scores (81% vs 13%), although physicians noted occasional omissions and hallucinations.	Patient Communication							
Zeinali 2024	Symptom-BERT: Enhancing Cancer Symptom Detection in EHR Clinical Notes.	This study aimed to develop a specialized language model, Symptom-BERT, to detect and extract cancer symptoms from clinical notes.	A Bio-Clinical BERT model was pretrained on one million unlabeled clinical documents, then fine-tuned on 1,112 annotated notes to identify 13 cancer symptom groups. The model's performance was validated internally and on 180 synthetic notes generated by ChatGPT-4, and compared against other BERT implementations.	Symptom-BERT achieved strong internal performance (F1=0.933, AUC=0.929) and good external performance (F1=0.831, AUC=0.834), detecting physical symptoms more accurately than psychological ones.	Item Identification							
Zhang 2024	The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant.	This study evaluated the potential of LLMs to enhance patient care by identifying specific diseases within EHR data and assisting clinicians in diagnostic and decision-making processes.	ChatGPT, GPT-4, and LLaMA were tested on gold-labeled EHRs from the MIMIC-III database to identify various diseases, comparing their performance (F1-scores) to disease-specific traditional ML models.	GPT-4 achieved near-equivalent or better F1-scores (≥85%) than disease-specific ML models for COPD, CKD, and PBC, outperforming them by 4.23% in the PBC case, while ChatGPT and LLaMA trailed in most metrics. Few-shot prompts improved precision and specificity but lowered sensitivity, underscoring the trade-offs in using LLMs for patient identification.	Item Identification							