Supplementary Table 2: Text mining models in emergency medicine

Title	Author	Type	Year	No of records	Methods	Performance metrics
A large language model-based clinical decision support system for syncope recognition in the emergency department: A framework for clinical workflow integration	Levra et al.	NLP	2025	30,320	NLP with Italian and multilingual BERT	AUC: 0.94-0.98
A pre-trained language model for emergency department intervention prediction using routine physiological data and clinical narratives	Ting-Yun Huang, et al	NLP	2024	176,81	Retrospective observational study using electronic health records, structured and unstructured clinical data, NLP preprocessing, and ML techniques.	BioClinicalBERT AUROC: 0.9 (best performance). ML Performance: TCND improved accuracy; deep learning models outperformed traditional models.zang
Advancing Emergency Department Triage Prediction with ML to Optimize Triage for Abdominal Pain Surgery Patients	Chai,et al	NLP	2024	38,214	Retrospective study using 9 years of ED triage data from electronic medical records (EMRs) - ML Models Tested: LightGBM, XGBoost, DNN, RF, LR - Training Data: 80% train, 20% test split - Feature Selection: Included structured data (vital signs, gender, age) and unstructured free-text data from triage notes	- LightGBM AUC: 0.899 (95% CI 0.891–0.903) - XGBoost AUC: 0.896 (95% CI 0.889–0.902) - DNN AUC: 0.896 (95% CI 0.888–0.902) - RF AUC: 0.889 (95% CI 0.880–0.896) - LR AUC (Baseline Model): 0.885 (95% CI 0.876–0.891) - Highest Accuracy: LightGBM (91.8%) - XGBoost had the highest net benefit in decision curve analysis
AI in the ED: Assessing the Efficacy of LLM Models vs. Physicians in Medical Score Calculation	Haim et al.	LLM	2024	150	AI-generated scores compared to actual patient outcomes	Human physicians achieved higher ROC-AUC on 3 out of 5 scores; AI models tended to be more cautious, leading to potential overtriage
An Ensemble Model for Predicting Dispositions of Emergency Department Patients	Kuo et al.	NLP	2024	80,073	Ensemble learning, BOW, TF-IDF	AUC: 0.94
Analyzing Pain Patterns in the Emergency Department: Leveraging Clinical Text Deep Learning Models for Real-World Insights	Hughes et al.	NLP	2024	235,789	Transformer-based deep learning, NLP, Interrupted Time Series Analysis	Accuracy: 93.4%, Pain incidence: 55.16%
Assessing the Precision of Artificial Intelligence in ED Triage Decisions: Insights from a Study with ChatGPT	Paslı et al	LLM	2024	758	Prospective observational study; GPT-4 was trained on local triage rules and compared to human triage decisions	High accuracy and specificity observed across different triage zones; minor discrepancies in yellow zone cases
Assessing the	Liu et al.	LLM	2024	45,00	Consistency: 93.3–	

Utility of Artificial Intelligence Throughout the Triage Outpatients: A Prospective Randomized Controlled Clinical Study					100%, Expert rating: 17/30 responses received 9.5-10 scores	
ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis	Hoppe et al	LLM	2024	100,00	Retrospective study, blinded assessment using a point-based grading system (0-2 scale)	GPT-4: 1.76/2, GPT-3.5: 1.51/2, Resident Physicians: 1.59/2
Developing and Evaluating LLM–Generated Emergency Medicine Handoff Notes	Hartman et al.	LLM	2024	1600	LLM customization and evaluation	ROUGE, BERTScore, SCALE
Early Identification of Suspected Serious Infection Among Patients Afebrile at Initial Presentation Using Neural Network Models and NLP	Choi et al.	NLP	2024	188,37	Retrospective study, NLP preprocessing using TF-IDF, artificial neural network-based modeling, external validation	AUC for Model 4 (best): Internal validation: 0.911 (95% CI: 0.906–0.915), External validation: 0.913 (95% CI: 0.909–0.917)
Emergency Department Triaging Using ChatGPT Based on ESI Principles	Colakca et al.	NLP	2024	745	Cross-sectional study	Accuracy: 76.6%; Kappa: 0.828
Enhancing Patient Safety in Prehospital Environment: Analyzing Patient Perspectives on Non-Transport Decisions With NLP and ML	Fahrat et al	NLP	2024	210,00	Sentiment analysis, topic modeling (LDA), supervised ML models (Naïve Bayes, SVM, RF, KNN)	Naïve Bayes and SVM models had an accuracy of 81.58% in predicting patient postrefusal actions
Estimation of racial and language disparities in pediatric emergency department triage using statistical modeling and natural language processing	Lee et al	NLP	2024	135,39	NLP-based clustering of chief complaints, KNN, MARS regression, LR models	African American children had 2.4–2.7x higher odds of low-acuity triage assignment; Non-English speakers also had higher odds of under-triage
Evaluating LLM-Assisted Emergency Triage: A Comparison of Acuity Assessments by GPT-4 and Medical Experts.	Haim et al.	LLM	2024	100	Observational study	Accuracy: Variability observed
Evaluating LLM-Based Generative AI Tools in Emergency Triage: A Comparative Study of ChatGPT Plus, Copilot Pro, and Triage Nurses	Arslan et al	LLM	2025	468,00	Prospective observational study, triage level assignment using Emergency Severity Index (ESI)	Triage accuracy: ChatGPT (66.5%), Copilot (61.8%), Nurses (65.2%); High-acuity detection: ChatGPT (87.8%), Copilot (85.7%), Nurses (32.7%)

Evaluating the accuracy of a state-of-the-art LLM for prediction of admissions from the emergency room	Gliksberg et al	LLM	2024	864,089	ML models: Bio-Clinical-BERT for unstructured triage notes, XGBoost for structured tabular data GPT-4 Experiments: Zero-shot, Few-shot, RAG, ML-informed predictions - Data Source: Electronic Health Records (EHRs) from 7 hospitals	- Best ML model (Ensemble): AUC = 0.878, AUPRC = 0.719, Accuracy = 82.9% - Naïve GPT-4: AUC = 0.79, AUPRC = 0.48, Accuracy = 77.5% - RAG Few-shot GPT-4: AUC = 0.821, AUPRC = 0.563, Accuracy = 81.3% - Best GPT-4 setup (RAG Few-shot + ML probabilities): AUC = 0.874, AUPRC = 0.709, Accuracy = 83.1%.
Evaluating the Reliability of ChatGPT as a Tool for Imaging Test Referral: A Comparative Study With a Clinical Decision Support System	Rosen et al	LLM	2024	97,00	Comparative study using ESR iGuide as a reference, evaluation of agreement levels, and subgroup analysis by specialists.	ChatGPT agreement with ESR iGuide: 87.6%; expert validation of ChatGPT's CAP CT recommendations: mean score 6.02/7
Evaluation of GPT-4 Ability to Identify and Generate Patient Instructions for Actionable Incidental Radiology Findings	Woo et al	LLM	2024	430,00	Retrospective analysis, NLP-based classification, and evaluation of AI-generated instructions	DA/PA-CC classification: Recall 99.3%, Precision 73.6%, F1-score 84.5%; DA-only classification: Recall 95.2%, Precision 77.3%, F1-score 85.3%
Identifying Incarceration Status in the EHR Using LLMs in Emergency Department Settings	Huang et al.	NLP	2024	1000	the classic BERT-based model and Clinical-Longformer model	Sensitivity: 100% (GPT-4); F1: 0.93 (Longformer)
Identifying Signs and Symptoms of UTI from ED Notes Using LLMs	Iscoe et al.	NLP	2024	1,250	NLP symptom extraction	F1 Score: 0.98 (Longformer); 0.96 (SpaCy)
Integrating Structured and Unstructured Data for Predicting Emergency Severity	Xingyu Zhang et al.	NLP	2024	8,716	Retrospective study	AUC: 0.789; accuracy: 0.726
LLMs Improve Identification of ED Visits for Symptomatic Kidney Stones	Bejan et al.	LLM	2024	500	GPT-4, GPT3.5, Llama-2 Logistic regression (LR), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM)	Best-F1: 0.833 (GPT-4)
Moving Biosurveillance Beyond Coded Data Using AI for Symptom Detection From Physician Notes	McMurry et al.	NLP	2024	85,678	NLP (NLP), AI-based symptom detection	F1-score: 0.796 (NLP) vs. 0.451 (ICD-10), Sensitivity: 0.93 (NLP) vs. 0.30 (ICD-10)
Near Real-Time Syndromic Surveillance of Emergency Department Triage Texts Using NLP: Case Study in Febrile Convulsion Detection	Khademi et al	NLP	2024	76,274	- NLP methods: Pattern matching, Standard ML (XGBoost), Deep Learning (BiGRU, CNN-BiGRU), Transformers (RoBERTa) - Data Augmentation: Synthetic text generation using LLM-2 - Data Source: 76,274 ED triage texts from	- RoBERTa (Transformer) F1-score = 0.921, Recall = 0.972, Precision = 0.875 - BiGRU F1-score = 0.900 - CNN-BiGRU F1-score = 0.899 - XGBoost F1-score = 0.822 - Pattern Matching F1-score = 0.797

					public hospitals in Victoria, Australia	
Novel Approach to Personalized Physician Recommendations Using Semantic Features and Response Metrics	Zheng et al.	NLP	2024	646,383	Semantic Features, Response Metrics, SBERT, Doc2Vec	F1-score: 76.25%, Service Quality: 41.05%, Rating: 3.90
Performance of ChatGPT on Prehospital Acute Ischemic Stroke and Large Vessel Occlusion Stroke Screening	Wang et al.	LLM	2024	400	LLMs, AI-driven stroke screening	AIS AUC: 0.75 (ChatGPT-4) vs. 0.59 (GPT-3.5), LVO AUC: 0.71 (ChatGPT-4) vs. 0.60 (GPT-3.5)
Performance of ML Versus the National Early Warning Score for Predicting Patient Deterioration Risk	Watson et al.	NLP	2024	174,393	ML, NLP, Triage text data	Average Precision: 0.92 (ML), 0.12 (NEWS)
Prediction of Hospitalization and Waiting Time Within 24 Hours of Emergency Department Patients	Seo et al.	NLP	2024	49,266	ML, NLP, AI-driven triage	AUC: 0.922, MAE: ~3 hours
Prediction of Outcomes After Cardiac Arrest by a Generative Artificial Intelligence Model	Amacher et al.	LLM	2024	713	Generative AI, Prognostic modeling	AUC: 0.85 (ChatGPT-4), OHCA (0.82), CAHP (0.83)
Racial Differences in Stigmatizing and Positive Language in Emergency Medicine Notes	Boley et al	NLP	2024	26,363	Transformer-based NLP model to classify language themes – LR models adjusted for demographics, insurance, comorbidities, and visit characteristics	- NLP Model F1-score: ≥0.89 for all theme classifications - Odds Ratios (ORs) for stigmatizing themes (NH Black vs. NH White patients): - Non-compliant: 1.26 ($p < 0.001$) - Financial difficulty: 1.14 ($p = 0.004$) - Skepticism: 0.87 ($p = 0.004$) - Odds Ratios (ORs) for NH Native American/AI patients: - Any negative theme: 2.02 ($p < 0.001$) - Substance abuse: 2.74 ($p < 0.001$) - Financial difficulty: 2.03 ($p < 0.001$)
Traditional ML, Deep Learning, and BERT (LLM) Approaches for Predicting Hospitalizations From Nurse Triage Notes: Comparative Evaluation of Resource Management	Patel et al	NLP	2024	1,391,988	- ML Models: BOW-LR-TF-IDF, XGBoost, Bi-LSTM (W2V) - Deep Learning Models: Bio-Clinical-BERT (Transformer-based) - Data Preprocessing: NLP tokenization, feature extraction - Training: 4 hospitals' data, external validation on a 5th hospital dataset	- Bio-Clinical-BERT AUC: 0.82 (10k records), 0.84 (100k records), 0.85 (1M records) - BOW-LR-TF-IDF AUC: 0.81 (10k), 0.83 (100k), 0.84 (1M) - XGBoost AUC: 0.76–0.82 - Bi-LSTM AUC: 0.78–0.84 - BERT had the highest sensitivity (0.81) but lower specificity (0.74)
Use of a LLM for Ambulance Dispatch and Triage	Shekar et al	LLM	2025	392	Retrospective comparative analysis of AI vs. human triage - Data Source: Real-world EMS dispatch requests randomized into 98 groups of four - LLM Prompting: Asked to identify highest-priority request - Validation: Three-person paramedic panel	- Overall Agreement: 76.5% LLM vs. Paramedics - Unanimous Panel Agreement: 93.8% LLM match rate - Majority Panel Agreement: 68.2% LLM match rate - Triage Uncertainty (Panel Disagreement): LLM struggled to determine priority

					reviewed same cases and voted on priority	
Use of a LLM to Assess Clinical Acuity of Adults in the Emergency Department	Williams et al.	LLM	2024	10,000	LLM-based triage, Emergency Severity Index (ESI) classification	Accuracy: 89% (LLM), 86% (Physician)
Using LLMs to Extract Core Injury Information from Emergency Department Notes	Choi et al.	LLM	2024	68,578	LLM-based information extraction, NLP	Accuracy: 0.935 (severity), 0.972 (intent)
Using ML and NLP in Triage for Prediction of Clinical Disposition in the Emergency Department	Chang et al.	NLP	2024	172,101	ML, NLP	Brier Score: 0.072-0.095 (internal/external validation)
Words to Live By: Using Medic Impressions to Identify the Need for Prehospital Lifesaving Interventions	Weidman et al	NLP	2025	12,913	NLP, GB ML	AUC: 0.793, Average Precision: 0.670
Exploring the potential of artificial intelligence models for triage in the emergency department.	Tortum F, et al.	LLM	2024	500	ChatGPT, Gemini, Pi AI	Triage exact agreement rates: ChatGPT (48.4%) > Gemini (44.7%) > Pi AI (28.7%) - Undertriage rates: ChatGPT (41.6%) > Pi AI (14.6%) > Gemini (7.8%) Nurses' performance was significantly better than AI models
Combination of ML algorithms with NLP may increase the probability of bacteremia detection in the emergency department	Haim G, et al.	NLP	2024	94,482	XGBoost, LR, NLP, free-text medical data analysis	AUC (75.6%), NLP contribution (4% improvement over tabular data alone)
Harnessing the Power of ML and Electronic Health Records to Support Child Abuse and Neglect Identification in Emergency Department Settings.	Landau AY, et al.	NLP	2024	33,963	XGBoost, NLP	Precision: 0.95, Recall: 0.88
ML-Based Prediction of Stroke in Emergency Departments.	Abedi V, et al.	NLP	2024	56,452	ML-based predictive modeling - XGBoost, RF, SVM, Generalized Linear Models - NLP applied to provider notes using Apache cTAKES - SHAP-based feature selection & dimensionality reduction via LSI	AUROC (structured EHR models): 0.88–0.92 AUROC (provider note-based models): 0.93–0.99 Sensitivity: 90% Specificity: 99% NPV: 87–90%, PPV: 80–98%
Early prediction of intensive care unit admission in emergency department patients using ML.	Pandey D, et al.	NLP	2024	484,094	NLP + GB model	- AUROC: 0.921 (30 min), 0.933 (240 min) - Accuracy: 92.6% (30 min) - Sensitivity: 72.5% (30 min), 74.1% (240 min) - Daily estimated ICU triggers: 2.7 per day

ML in clinical practice: Evaluation of an artificial intelligence tool after implementation.	Akhlaghi H; et al	NLP	2024	7,125	NLP of triage notes - LR and ensemble learning models trained on 10-year historical data	- AUROC: 0.74 (real-time model) vs. 0.83 (pre-implementation model) - Sensitivity: 73.1% - Specificity: 74.3% - F1-score: 0.59