Review

AI-Enabled Text Mining: A Paradigm Shift in Disease Prediction, Drug Discovery, and Clinical Research

Jayakumar Manoharan[1] | Yamini Sehgal[1] |

1. Electric Power Research Institute, 1300 W W T Harris Blvd, Charlotte, NC, USA – 28262.

Abstract

The exponential growth of unstructured biomedical data, comprising nearly 80% of healthcare information, has created both unprecedented opportunities and formidable challenges for extracting actionable insights from clinical notes, electronic health records, biomedical literature, and patient-reported outcomes. Traditional computational methods are inadequate to handle the massive scale, semantic complexity, and heterogeneity of such data, necessitating advanced artificial intelligence (AI)-enabled text mining approaches. This review examines the evolution of AI-driven text mining from experimental application to clinical necessity, with emphasis on disease prediction, drug discovery, and clinical research. Literature encompassing natural language processing (NLP), transformer-based models, clinical case studies, and regulatory frameworks was systematically analyzed across oncology, cardiology, neurology, and pharmacovigilance domains. AI-enabled text mining demonstrates robust performance across multiple applications: disease prediction models achieve 67–98% accuracy in early diagnosis and risk stratification; transformer-based methods yield 80.6% F1-scores for drug–target interaction extraction; and adverse drug reaction detection from social media achieves 84.2% sensitivity and 98% specificity. In clinical research, systematic review timelines are reduced by up to 70%, and clinical trial recruitment screening requirements decline by nearly 80%. Real-time clinical decision support powered by large language models reduces diagnostic time from over 30 minutes to less than one minute, maintaining accuracy comparable to expert teams. Despite remarkable progress, challenges persist, including data heterogeneity, annotation quality, computational demands, translational gaps, algorithmic bias, and privacy concerns. Future directions include multimodal integration with genomics, imaging, and biosensors, explainable AI frameworks, and federated learning for collaborative research. AI-enabled text mining represents a transformative paradigm shift toward predictive, preventive, and personalized medicine, bridging the gap between exponential data growth and human cognitive limitations while improving patient outcomes and accelerating scientific discovery.

ISSN: 3108-2696 (Online)

1. Introduction

Over the past two decades, the world has witnessed an extraordinary digital transformation across pharmacological sciences, encompassing healthcare, pharmaceutical innovation, and biomedical research. This transformation has not only reshaped how medical knowledge is generated, stored, and utilized but has also accelerated the integration of data-driven approaches into every facet of modern medicine. The global healthcare analytics market exemplifies this explosive growth trajectory, expanding from USD 55.18 billion in 2024 to a projected USD 371.49 billion by 2033, marking a remarkable compound annual growth rate (CAGR) of 23.6% **(Almeman 2024)**. This exponential growth reflects not merely technological progress, but a fundamental paradigm shifts towards data-centric medicine, where each patient encounter, research initiative, and clinical observation contributes to an ever-expanding reservoir of biomedical intelligence. Healthcare systems now generate data at an unprecedented scale, with the industry contributing approximately 30% of the world's total data volume **(Batko and Slezak 2022)**. The compound annual growth rate of healthcare data is projected to reach 36% by 2025, creating an information ecosystem of staggering complexity. To illustrate this scale, PubMed alone hosts over 36.5 million indexed citations as of 2023, with approximately 1.5 million new publications added annually. Additionally, PubMed Central maintains over 11.2 million full-text articles, while clinical documentation, electronic health records, and social media platforms generate tens of petabytes of unstructured text annually, creating an immense knowledge repository within which hidden patterns reveal disease trajectories, adverse drug reactions, population health trends, and epidemiological transitions with profound implications for patient care and drug discovery. A pivotal dimension of this transformation lies increasingly within unstructured textual data, which accounts for approximately 80% of healthcare information. Beyond structured datasets such as diagnostic codes, laboratory values, and imaging parameters, healthcare systems are now confronted with immense textual ecosystems. This consisting of electronic health records (EHRs), patient histories, physician notes, biomedical literature, -regulatory documents, and real-world evidence derived from public data sources and patient forums **(El-Warrak *et al.* 2023)**. This unstructured data includes medical imaging data, which constitutes 80% of all clinical content, clinical notes that capture the severity and nuances of patient conditions, and free-form documentation that often provides insights far beyond what structured diagnostic codes can convey. Yet, the sheer scale and complexity of this information present formidable challenges. Biomedical language is inherently complex, characterized by semantic ambiguities, contextual variations, and discipline-specific terminologies that render conventional computational methods or human-driven analysis inadequate. Clinical notes frequently display stylistic and linguistic heterogeneity across specialties, while biomedical publications grow at a rate far exceeding the capacity of human cognition **(Awrahman *et al.* 2022)**. Health-related discussions on social media, although rich in real-world context, are often embedded within noisy, unstructured conversations shaped by colloquial expression. The inevitable consequence is a widening gap between the prolific generation of biomedical big data and its meaningful translation into actionable knowledge **(Kanchan and Gaidhane 2023)**. This review examines the use of AI-driven text mining in the biomedical field, encompassing electronic health records (EHRs), clinical documentation, literature, and patient stories, with a focus on its applications in predicting diseases, discovering

Received on	:	2025-10-10	**Key Words:**
Revised on	:	2025-11-27	Artificial intelligence
Accepted on	:	2026-01-07	Text mining
Published Online	:	2026-04-28	Natural language processing
Review Model	:	Single-Blind Review	Data Mining
No. of Reviewers	:	Two	Machine learning
Edited by	:	Dr Chandrabose Selvaraj	Healthcare
Vol and Issue	:	02 (02)	Innovation
Page No	:	23-35	
Plagiarism Level	:	14% and 00% (AI)	DOI: 10.64659/jomi/215260
Correspondence	:	Dr. M. Jayakumar	
Contact Author	:	✉	

drugs, and conducting clinical research. The discussion excludes non-textual AI techniques like imaging or speech analysis. In this context, "text mining" specifically pertains to natural language processing (NLP) methods utilized to analyse and understand unstructured medical text. Artificial intelligence (AI) and text mining have emerged as decisive solutions to bridge this critical gap. Advances in machine learning, deep learning, and natural language processing (NLP) now enable the adaptive capture of semantic richness and contextual depth beyond the reach of keyword searches or rule-based computational systems. The global NLP market in healthcare is projected to grow from USD 2.7 billion in 2023 to USD 11.8 billion by 2028, representing a CAGR of 34.4%. Large language models (LLMs), hybrid AI architectures, and transformer-based frameworks such as BioBERT, PubMedBERT, and BioGPT have demonstrated unprecedented performance in structuring unstructured data, transforming it into queryable, knowledge-rich assets that democratize access to biomedical intelligence and accelerate evidence-informed decision-making across healthcare systems worldwide **(Yang *et al.* 2023)**.

The clinical and translational applications of AI-enabled text mining are vast and measurably impactful. In disease prediction, advanced NLP models can analyse patterns within EHRs and clinical documentation with diagnostic accuracy improvements of 20-30% in cardiovascular and diabetic cases, while AI-based clinical decision support systems have demonstrated accuracy rates ranging from 81% to 99.7% across various diagnostic applications **(Lee *et al.* 2024)**. In drug discovery, AI has the transformative potential to reduce development timelines from the traditional 10-15 years to approximately 7-9 years, with some processes that typically take several years being compressed to mere months. AI-powered virtual screening and compound identification can cut R&D timelines by up to 50%, while machine learning algorithms can shave up to four years off the total drug development timeline **(Dermawan and Alotaiq 2025)**. Clinical research has equally benefited from these technological advances. AI-powered systematic review tools can eliminate 20-88% of titles and abstracts from human review, providing time savings of 7-86 hours per review. These automation tools demonstrate sensitivity rates ranging from 79% to 98.75% and specificity rates from 82% to 100%, though their implementation requires careful consideration of accuracy-efficiency trade-offs. The technology enables researchers to process vast biomedical literature repositories, mine compound databases, and analyse clinical trial data with unprecedented efficiency, effectively reducing systematic review redundancy by up to 40% while enhancing methodological rigor and reproducibility **(van Mossel *et al.* 2025)**.

Moreover, real-world evidence derived from patient records, social media discourse, and clinical documentation provides deeper insights into therapeutic outcomes, adverse events, and population health dynamics. NLP techniques can capture 50% more clinical cases than structured data alone, enabling more comprehensive patient monitoring and predictive analytics capabilities that transform reactive healthcare into proactive, precision-based interventions **(Murray *et al.* 2024)**. This review critically explores how AI-enabled text mining has evolved from experimental technology into a clinical necessity, representing a cornerstone of modern biomedical knowledge management. The global big data in healthcare market is estimated to grow from USD 78 billion in 2024 to USD 540 billion by 2035, with a CAGR of 19.20%, underscoring the massive investment and strategic importance of these technologies **(Wang *et al.* 2025)**. Methodological breakthroughs now allow machines to analyze and structure clinical language with human-level precision, facilitate the integration of heterogeneous textual datasets, and directly inform workflows in disease prediction, research optimization, and drug discovery. At the same time, the rapid expansion of these technologies underscores significant challenges, including algorithmic bias, interpretability, data privacy, regulatory compliance, and generalizability across diverse populations and healthcare systems **(Vamathevan *et al.* 2019)**. Current AI tools in systematic reviews show varying levels of accuracy and efficiency, with 95% of pharmaceutical companies now actively investing in AI capabilities, yet implementation often requires semi-automation approaches that balance accuracy with operational efficiency. By synthesizing representative case studies, translational applications, and performance evaluations, this review positions AI-enabled text mining not merely as a computational tool but as a transformative paradigm in medicine **(Ge *et al.* 2024)**.

The integration of text mining with AI offers a viable pathway to address one of the most pressing challenges of 21st-century healthcare: transforming the vast biomedical data landscape into actionable intelligence. As the global health system increasingly pivots towards value-based care and precision medicine frameworks, the scalable deployment of AI-driven text mining will determine institutional competitiveness, the pace of research innovation, and ultimately, patient outcomes **(Chong *et al.* 2025)**. This review provides both a comprehensive analysis of existing capabilities and a strategic perspective on leveraging this technological revolution in shaping the future of predictive, preventive, and personalized medicine. More than technical documentation, it envisions a new era where artificial intelligence becomes the critical interface between human expertise and biomedical big data, enhancing clinical decision-making and accelerating scientific discovery in the age of digital transformation. The evidence demonstrates that we stand at an inflection point where AI-enabled text mining is transitioning from promising technology to essential infrastructure, fundamentally redefining how biomedical knowledge is discovered, validated, and translated into improved patient care **(Al Kuwaiti *et al.* 2023)**.

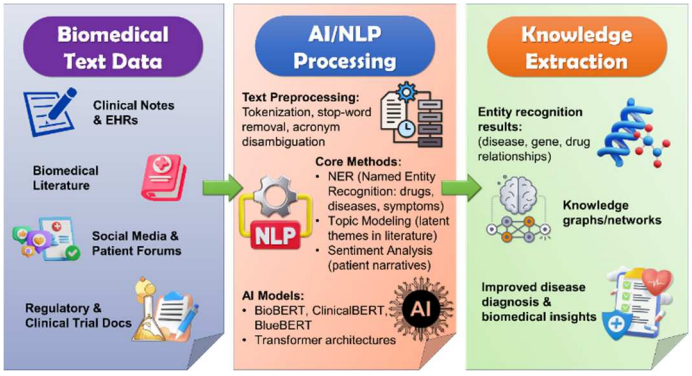

Figure 1: Fundamentals of AI-enabled Text Mining in the Medical Domain

2. Fundamentals of Text Mining in the Medical Domain

In medical domain, text mining is one of the sophisticated convergences of computational linguistics and healthcare informatics, which employing specialized natural language processing (NLP) applications to extract meaningful information form biomedical text. Entity recognition (NER) is the core methodology of text mining for identifying medical entities including, drugs, disease, and symptoms; topic exhibiting for identify the latent thematic structures in clinical literature; and sentiment analysis for enumerating demonstrative and evaluative content in clinical narratives. These basic and effective approaches achieve significant accuracy, with stat-of-the-art medical NER system achieving 89.8% precision in clinical entity extraction tasks **(Ahmad *et al.* 2023)**. AI and machine learning incorporation has resulted in biomedical text mining via transformer-based structures, specifically BERT variants optimized for medical domains. BioBERT is effectively used in PubMed abstract and PMC full-text articles, demonstrating the effectiveness of with 87.6% of F1 score compared to general BERT's 81.0% F1 score on medical NER tasks. Other variants such as clinical BERT, and BlueBERT achieves 83.5% and 85.0% of F1 score respectively, which highlights the domain-specific

optimization benefits. Recent achievements include Bioformer, a compact model delivering 99% of PubMedBERT's performance with 40% of parameters to achieve 2-3-fold faster inference speed. Large language model like ChatGPT shows promising ability in biomedical text generation and understanding, however, the performance differs across specific clinical tasks **(Lee *et al.* 2020)**.

Processing biomedical languages represents the unique challenges curtailing from inherent linguistic complexity. Acronym disambiguation represents a critical bottleneck, with clinical texts comprising 24,090 abbreviations having several meanings, the abbreviation "PA" alone possesses 142 possible elucidations including pancreatic adenocarcinoma, physician assistant, and arterial pressure. Automated disambiguation systems attain 87.9% accuracy using progressive machine learning approaches; however, performance degrades with rare abbreviations **(Moon *et al.* 2012)**. Sentiment analysis in clinical narratives faces other complexity, with accuracy ranging from 71.5-88.2% owing to objective medical language that shortages clear emotional markers typical in general domain texts. But in case of dealing with multiple languages in medical text, significantly increase the challenges specifically in non-English biomedical literature, where medical terms can vary a lot across culture contexts. To address this, incorporating the domain-specific vocabularies and contextual embedding and cross lingual transfer learning are emerging as effective approaches to overcome issues including basic linguistic barriers in medical text mining applications **(Figure 1) (Denecke and Reichenpfader 2023)**. Ontology alignment and terminology normalization assist in standardizing medical terminology by linking synonyms and abbreviations to a unified concept through resources such as UMLS, SNOMED CT, and MeSH. This minimizes confusion and enhances the precision of biomedical text mining models. BioBERT and PubMedBERT are based on the BERT architecture; however, BioBERT is pretrained using texts from both PubMed and PMC, while PubMedBERT is solely trained from scratch on PubMed, resulting in improved alignment with biomedical vocabulary. BioGPT, a generative model inspired by GPT, is trained on PubMed articles, which enhances its capabilities for text generation and summarization. In practice, BioBERT and PubMedBERT are commonly utilized for extracting clinical texts, whereas BioGPT is employed for creating clinical summaries and providing insights from the literature.

3. Data Sources and Repositories for Medical Text Mining

Medical text mining influences an unprecedented environment of diverse data repositories that collectively signify the world's largest biomedical information infrastructure. These sources incorporate structured clinical data, peer-reviewed literature, regulatory databases, and emerging real-world evidence platforms, each contributing exclusive viewpoints to inclusive healthcare analytics **(Gonzalez *et al.* 2013)**. Electronic Health Records (EHRs) establish the keystone of clinical text mining, with over 90% of US hospitals having adopted EHR systems as of 2021. The global EHR market, valued at $29 billion in 2020, is projected to reach $47 billion by 2027. EHRs contain approximately 1.28 million records per healthcare system, with studies demonstrating successful data mining applications in rare disease identification, cohort discovery, and biomarker analysis. Modern EHR systems generate petabytes of unstructured clinical notes annually, including physician narratives, nursing documentation, and patient correspondence. Text mining algorithms achieve 97% sensitivity in identifying adverse drug reactions from EHR free-text notes, with positive predictive values reaching 70% using advanced NLP preprocessing. However, EHR text mining faces challenges including data heterogeneity, missing information, and inconsistent documentation practices across healthcare institutions **(Modi and Feldman 2022)**. The text from electronic health records needs to undergo various preprocessing techniques, such as de-identification, cleaning, tokenization, expanding abbreviations, standardizing medical terminology, and eliminating unnecessary characters or noise. These procedures assist in organizing the unrefined clinical text and enhance the effectiveness of subsequent NLP models. Clinical trial databases, exemplified by ClinicalTrials.gov, represent the world's largest clinical research repository containing information on over 400,000 studies from 220 countries. The database receives approximately 330 new registrations and 2,000 revisions weekly, with 30 new results submissions processed continuously **(Zarin *et al.* 2011)**.

ClinicalTrials.gov's structured data architecture includes participant flow information, baseline characteristics, outcome measures, and adverse events in standardized tabular formats. Advanced API capabilities enable programmatic access to trial data, facilitating large-scale text mining of protocol descriptions, eligibility criteria, and outcomes reporting. Recent studies demonstrate successful extraction of drug-indication relationships and safety profiles from clinical trial narratives, though challenges persist in standardizing outcome measure descriptions and ensuring data completeness **(Tse *et al.* 2009)**. Biomedical literature repositories, primarily PubMed/MEDLINE, contain over 35 million indexed citations with continuous expansion including preprints from bioRxiv and medRxiv. Advanced text mining tools like BioTextQuest v2.0 can process up to 10,000 PubMed abstracts simultaneously, employing machine learning algorithms for document clustering and entity recognition. The platform utilizes EXTRACT named entity recognition, identifying genes, proteins, chemicals, diseases, and Gene Ontology terms with high precision. Systematic text mining applications include drug discovery, biomarker identification, and systematic review automation, with tools achieving significant time reduction in literature analysis compared to manual approaches. Modern pipelines integrate TopicTracker and semantic analysis to provide comprehensive literature mining from querying to visualization **(Theodosiou *et al.* 2024)**.

Social media and patient-reported outcomes represent rapidly growing data sources for real-world evidence (RWE) generation. Analysis of 367,573 patient stories from platforms like Care Opinion reveals comprehensive healthcare experiences spanning communication quality, clinical services, and patient satisfaction. Social media text mining achieves 99% sentiment classification accuracy with over 55% of treatment-related posts expressing negative sentiment, indicating patient dissatisfaction. Pharmaceutical companies like Roche successfully employ NLP-based social media mining to analyze Parkinson's patient discussions across multiple platforms, providing insights into patient-reported outcomes and competitive intelligence. Twitter-based pharmacovigilance studies process billions of tweets, with machine learning pipelines identifying adverse drug events and drug-drug interactions from patient discussions. However, regulatory acceptance remains limited due to challenges in establishing causal relationships and ensuring data verification **(Zakkar and Lizotte 2021)**.

Multi-omics databases and integration with text-based evidence represents the frontier of personalized medicine research. Major repositories including The Cancer Genome Atlas (TCGA), Alzheimer's Disease Neuroimaging Initiative (ADNI), and Genotype-Tissue Expression (GTEx) provide comprehensive multi-modal datasets combining genomics, transcriptomics, and metabolomics with clinical narratives. Integration methodologies encompass simultaneous and stepwise approaches, with unsupervised learning techniques enabling disease subtyping, biomarker discovery, and pathway analysis **(Yang *et al.* 2025)**. Advanced platforms like GraphOmics facilitate multi-omics data exploration and hypothesis generation through correlation analysis and network visualization. Text mining integration with omics data enables identification of novel biomarkers, drug targets, and therapeutic mechanisms by correlating literature findings with molecular signatures. Semantic technologies enhance data standardization and analysis across

heterogeneous omics datasets, though computational complexity and interpretation challenges persist. These integrated approaches demonstrate promise in cancer research, neurodegenerative diseases, and precision medicine applications were molecular data interpretation benefits from contextual literature knowledge **(Figure 2)** **(Wandy and Daly 2021)**.

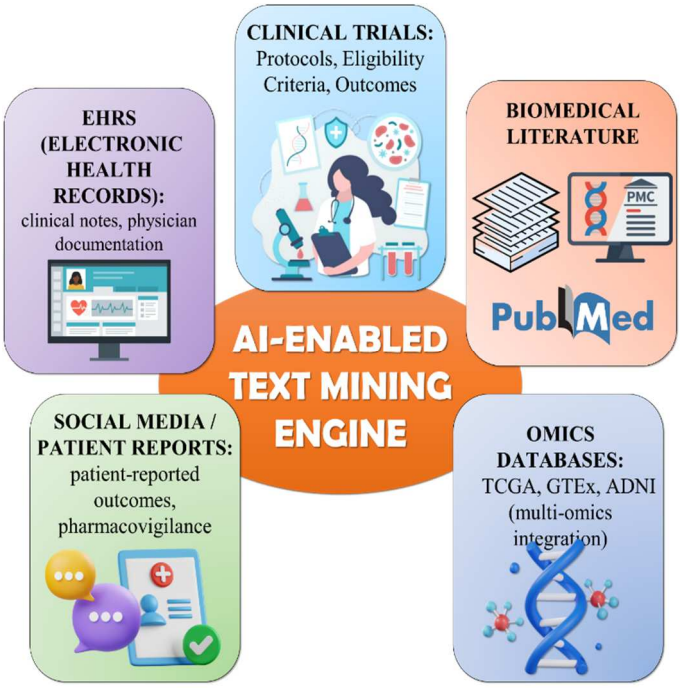

Figure 2: Data Sources and Repositories in AI-Enabled Medical Text Mining

4. AI-Enabled Text Mining in Disease Prediction

Early diagnosis and risk stratification from EHRs have achieved remarkable breakthroughs through AI-enabled text mining, demonstrating significant clinical impact. Advanced hybrid models combining textual and numerical EHR data achieve 67.0% precision@10 in early-stage disease prediction, substantially outperforming single-modality approaches **(Dai *et al.* 2025)**. Machine learning algorithms analyzing clinical narratives demonstrate superior performance in cardiovascular risk prediction, with Random Forest and deep learning models achieving AUCs of 0.865 and 0.847 respectively, significantly outperforming conventional risk scores (AUC 0.765). Deep learning approaches utilizing longitudinal EHR data predict Type 2 diabetes, COPD, hypertension, and myocardial infarction with AUCs ranging from 0.92-0.94 over 3-year prediction horizons **(Hossain *et al.* 2024)**. Text mining algorithms excel in adverse drug reaction detection from clinical notes, achieving 97% sensitivity with 70% positive predictive value. Comorbidity identification and disease trajectory modeling represent critical applications where AI-driven text mining provides unprecedented insights**(van de Burgt *et al.* 2024)**.

Machine learning analysis of COVID-19 patient data identifies hypertension, diabetes, chronic kidney disease, cardiovascular disease, and COPD as the most significant mortality predictors, achieving >80% accuracy across six different algorithms. XGBoost models demonstrate 11% improvement in C-statistics compared to traditional Cox regression models for cardiovascular risk prediction by incorporating longitudinal EHR patterns. Advanced deep learning frameworks achieve net reclassification improvement rates of 3.9% for cardiovascular events by analysing irregular, repeated measurements in real-world clinical data. Comorbidity prediction models utilizing machine learning achieve 96% accuracy in identifying high-risk patient populations across multiple disease domains **(Seyedtabib *et al.* 2024)**.

Outbreak prediction and epidemiological surveillance have been revolutionized through AI-enabled social media and clinical text mining. COVID-19 surveillance systems utilizing text mining of emergency care records detect outbreaks 15-72 days before traditional notification systems, with cross-correlations ranging from 0.82-0.93 across pandemic waves. Advanced text mining algorithms processing 2.76 million medical records achieve early outbreak detection through symptom pattern analysis, enabling proactive public health interventions **(Rocha *et al.* 2024)**.

Social media text mining using recurrent neural networks successfully identifies COVID-19 infected individuals from Twitter discussions, outperforming traditional machine learning algorithms for disease surveillance purposes. Google Trends-based machine learning models predict COVID-19 incidence with RMSE of 7.562, identifying handwashing and sanitizer searches as key predictive factors. ProMED-mail text mining using TextRank algorithms successfully tracks Ebola and Zika outbreak evolution, with keyword co-occurrence networks matching WHO epidemic timelines **(Cheng *et al.* 2025)**. Case studies demonstrate exceptional performance across medical specialties. In oncology, AI-driven text mining of pathological reports achieves 98% specificity and sensitivity in prostate cancer diagnosis. Radiomics-based AI models identify early-stage lung cancer with improved detection rates compared to traditional imaging methods. PathAI's algorithms analyzing histopathological images demonstrate superior accuracy to pathologists in breast cancer detection **(Zhang *et al.* 2023)**.

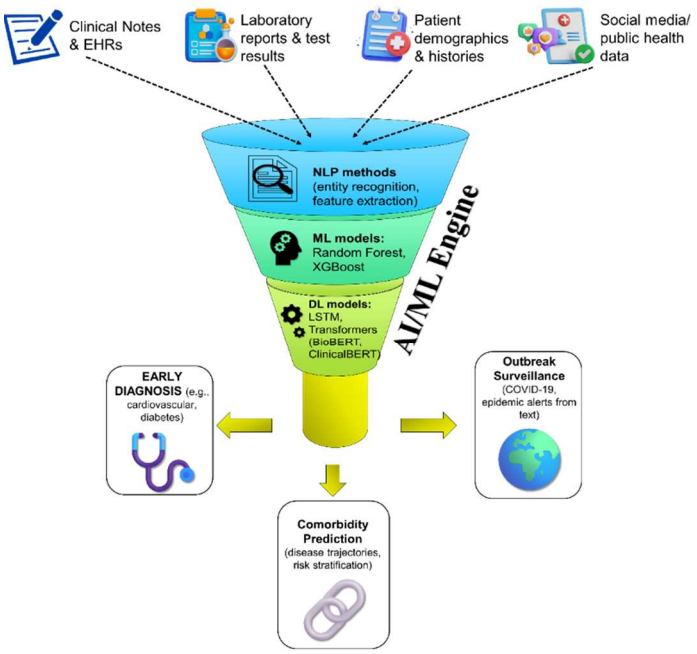

Figure 3: AI-Enabled Text Mining Pipeline for Disease Prediction

In cardiology, machine learning models utilizing EHR text data achieve 26.25% improvement in AUROC when augmenting neurologist assessments for cardiovascular risk stratification. CNN-LSTM hybrid frameworks reliably outperform traditional classifiers in heart disease prediction from clinical narratives **(Kumar *et al.* 2025)**. In neurology, AI-models achieve 0.94 micro averaged AUROC in differentiating dementia etiologies with 26.25% enhancement in diagnostic accuracy when expanding neurologist assessment. In the early detection of Alzheimer's diseases with AI models incorporating the hippocampal imaging and clinical text achieve robust performance across the wide range of patient populations. These applications significantly demonstrate the AI-enabled text mining's transformative potential in precision medicine, qualifying prior intercession, personalized treatment approaches and enhance the

patient outcomes across the disease spectrum **(Figure 3) (Xue *et al.* 2024)**. Social media contributes to early outbreak forecasting as individuals freely share their symptoms, travel experiences, and current local disease situations before official data is available. AI technologies gather posts by utilizing keywords, hashtags, geolocation, and public APIs. The information goes through a filtering process to eliminate spam, irrelevant material, bots, and duplicates using natural language processing for cleaning and classification. Following the filtering stage, machine learning algorithms identify abrupt increases in posts related to symptoms, enabling the detection of COVID-19 outbreaks several weeks sooner than conventional surveillance methods. While AI and NLP models demonstrate impressive results in research environments, their effectiveness in real-world scenarios hinges on the quality of data, how well they fit into clinical workflows, and the need for ongoing validation. In reality, accuracy frequently diminishes due to the variability and noise found in clinical text across different hospitals. Thus, although achieving high benchmark scores indicates promise, realizing actual benefits necessitates meticulous implementation and continuous improvement.

5. Case Studies in AI-Enabled Text Mining

Real-world applications of AI-enabled text mining across various healthcare domains expose their transformative capability through computable outcomes that encompass far beyond theoretical potential. These applications validate computable developments in clinical efficiency, patient safety, cost reduction, and physician security, while also prominence critical success factors and implementation challenges. For instance, Mass General Brigham pioneered revolutionary ambient documentation and crisis management through a calculated execution architecture that addressed both acute crisis response and chronic healthcare challenges **(Bajwa *et al.* 2021)**. During the COVID-19 pandemic, their AI-drove voice system not only achieved devastating patient volumes but also essentially reformed patient interaction workflows. In its first week alone, the chatbot processed more than 40,000 patient interactions, demonstrating AI's capacity to rapidly adapt to healthcare crises within existing infrastructure constraints **(Jadczyk *et al.* 2021)**.

The most striking outcome was a documented reduction in physician burnout, with a 21.2% absolute reduction at 84 days and a 30.7% increase in documentation-related wellbeing at 60 days in collaboration with Emory Healthcare. This represents an unprecedented intervention impact, as no other healthcare technology has shown comparable burnout mitigation at such scale. Financial and operational analysis further reveals that AI scribes reduce documentation costs by 60–75% compared to human alternatives, with practices saving $122,000 annually and physicians reporting reclaimed nights and weekends, even describing how they "rediscovered their joy of practicing medicine." The system also increased encounter volumes by 5%, adding $54,000 in annual revenue per user**(You *et al.* 2025)**. Scalability was equally remarkable, with the program growing from 18 physicians in July 2023 to more than 800 by July 2024, and over 3,000 by April 2025, a 16,600% expansion in under two years. Clinical quality also improved as 90% of clinicians were able to provide undivided attention to patients compared to 49% previously, raising patient satisfaction scores by 4.4 percentage points for "concern shown by provider," proving that augmentation enhances rather than diminishes human connection in care delivery **(Al-Abri and Al-Balushi 2014)**. At Mayo Clinic, AI-enabled text mining through IBM Watson transformed precision oncology and clinical trial enrolment. Traditionally, only 5% of cancer patients participate in clinical trials, yet Watson achieved an 80% increase in breast cancer enrolment over 11 months, a performance that would otherwise take years. Independent studies validated 87.6% overall eligibility accuracy for breast protocols and 74.9% for lung, reducing processing times from 7.4 hours to mere minutes. Attribute accuracy was also high, with 93% for HER2 status and 88.6% for hormone receptor status **(Zhou *et al.* 2019)**. International concordance analysis revealed geographic variation, such as 96% concordance for ovarian cancer in Chinese centers but only 12% for gastric cancer, highlighting cultural and practice dependencies. Patient outcomes were significantly improved, with higher response rates, longer progression-free survival, and better overall results. As Dr. Tufia Haddad explained, Watson enabled physicians to efficiently develop comprehensive treatment plans, embedding clinical trials into standard care. Watson expanded to lung and gastrointestinal cancers across more than 14 countries, including China, India, Korea, and Bangladesh, but later challenges and eventual divestiture emphasized the importance of transparency, workflow integration, and realistic communication of AI capabilities **(Algera *et al.* 2023)**.

Similarly, Mount Sinai Health System demonstrated predictive analytics excellence. Their delirium detection model increased detection by 400%, from 4.4% to 17.2%, across a study of 32,000 patients, addressing a condition that affects up to half of hospitalized patients and elevates mortality and cognitive decline. Their fall prediction model generated millions in potential savings by preventing incidents that cost $30,000 each, while their malnutrition detection system improved predictive value from 20% to 70%, a 250% gain with direct benefits for wound healing, infection reduction, and readmission prevention **(Friedman *et al.* 2025)**. Mount Sinai also reduced emergency department wait times by 30% using predictive modeling based on weather, events, and historical patterns, with a 50% improvement in patient flow efficiency. Their multimodal data integration approach drew on 14 AI applications using structured EHR data, clinical notes, images, and vital signs, creating a holistic system beyond human cognitive capacity. Cost efficiency innovations further showed that task-grouping strategies could reduce API costs up to 17-fold, making advanced AI sustainable at scale **(Shameer *et al.* 2017)**.

The University of Missouri Health Care, in partnership with Cerner, integrated AI into EHR intelligence for early sepsis detection using NEWS scoring, enabling life-saving interventions in a condition where timing is critical. By addressing the 49% of physician time consumed by EHR-related tasks, they improved productivity, job satisfaction, and patient interaction. Their collaborative innovation model, built through the Tiger Institute partnership, emphasized continuous clinical feedback rather than traditional procurement, ensuring relevance and adaptability **(Schinkel *et al.* 2023)**. Johns Hopkins Hospital extended predictive analytics globally with its ACG System, refined over 25 years. Laboratory incorporation enhanced prophecy accuracy, hovering R^2 by up to 3.7%, improving high-cost patient documentation by 121%, and augmenting inpatient admission prediction by 188%. The application of neural networks and regression-based methods to forecast COVID-19 cases up to 14 days prior across multiple countries, to understand the AI's global public health values **(Tan *et al.* 2023)**. Beyond the clinical system, HSBC applied AI-enabled text analysis in financial compliance scanning millions of communications daily with aid of NLP models. This system effectively finds the behavioural patterns suggestive of market abuse which helped accelerate the investigation and align more closely with regulatory standards. Collectively, these cases demonstrate the important success factors; ROI analysis signifies the returns exceeding 12,000% with payback in one to three months, however it requires clinical partnerships, careful expectation management and workflow integration to achieve such success. Conversely, IBM Watson for Oncology's $4 billion failure highlighted the pitfalls of synthetic data and poor workflow integration, inadequate transparency which proving the success depends on alignment between problems and solutions **(Bajwa *et al.* 2021)**.

Transversely all organisations, effective scalability contexts share traits such as real-world data use, governance structures, continuous refinement, and communicate clearly and realistically. Taken together,

these case studies demonstrate that AI-enabled text mining has attained clinical maturity with assessable enhancements in patient outcomes, operational efficiency, and physician wellbeing. The healthcare paradigm is shifting is no longer just toward predictive, personalized, and data-driven medicine, but as a demonstrable operational reality across top organization worldwide **(Maleki Varnosfaderani and Forouzanfar 2024)**. Ultimately, continued success depends less on technology itself than on strategic vision, rigorous validation, clinical partnerships, and continuous improvement.

6. AI and Text Mining in Drug Discovery and Repurposing

Literature mining for drug-target interactions has developed as a transformative method in computational drug discovery, accomplishing notable precision through advanced AI methodologies. Ensemble transformer models using both gene descriptions from Entrez Gene database and chemical descriptions from Comparative Toxicogenomics Database and achieve the highest performance on the DrugProt hidden test set with 80.6% F1scores among the other submitted models. These methods exploit pre-trained language models like BERT to extract drug-target interactions as entity-relationship problems, suggestively outperforming traditional methods. Transformer-based models demonstrate exceptional capability in extracting deep features from sequence-only information, with CoaDTI framework achieving significant performance improvements through co-attention mechanisms applied to protein sequences and drug SMILES strings. Graph neural networks integrated with sequence embeddings further enhance molecular representation learning, enabling comprehensive DTI prediction that combines multiple data modalities including chemical structures, protein sequences, and biological networks. Adverse drug reaction detection from patient records and social media has achieved clinical-grade accuracy through sophisticated NLP frameworks.

The ADR Detection Framework (ADF) utilizing MetaMap and UMLS achieves 84.2% sensitivity and 98% specificity in warfarin-related adverse event detection from social media discussions. Machine learning algorithms processing patient reviews from askapatient.com, webmd.com, and iodine.com successfully identify previously unreported ADRs, with multiple organ system effects being the most frequently reported at 1.50%. Advanced context-aware algorithms like aTarantula utilize FastText embeddings and aggregated lexicons to extract contextual data from patient forums, achieving high precision in detecting drug-specific adverse events. These systems demonstrate particular value in diabetes drug ADR detection, uncovering age and gender differences in adverse reactions while providing real-time pharmacovigilance capabilities. BERT-based ensemble models achieve state-of-the-art performance in social media ADR classification tasks, outperforming traditional machine learning approaches **(Aldahdooh *et al.* 2024)**.

Text-driven drug repurposing strategies leverage comprehensive literature mining and semantic inference to identify novel therapeutic applications. Network-based approaches like HeTDR combine heterogeneous networks with text mining to achieve superior performance in drug repositioning tasks. Statistical analysis methods applied to EHR data demonstrate significant efficacy, with chi-square tests and false discovery rates revealing that theophylline-treated patients show 0.11% glaucoma prevalence compared to 0.058% in celecoxib-treated patients, suggesting potential antiglaucoma effects. Text mining approaches utilizing biological ontologies enable cross-domain knowledge discovery, where nutritional deficiencies linked to diseases in one study can inform drug repurposing decisions for related conditions. Machine learning algorithms analyzing EHR cohorts achieve significant associations between drug exposure and disease outcomes through Cox regression and logistic regression models, enabling systematic identification of repurposing candidates. Integration with computational docking, network pharmacology, and AI-driven drug design represents the frontier of modern drug discovery **(Jin *et al.* 2021)**. Pathway-based repurposing approaches leverage metabolic and signaling pathway information to predict drug-disease connections, offering holistic perspectives that traditional reductionist methods might miss. Deep learning innovations including convolutional neural networks for molecular structure analysis, recurrent neural networks for sequential biological processes, and graph neural networks for complex system modeling demonstrate exceptional performance in predicting binding affinities and drug effects. The COVID-19 pandemic showcased these capabilities when deep learning methods successfully identified baricitinib as a potential treatment through AI-based screening, later validated in clinical trials. Multi-omics integration with text-based evidence enables comprehensive drug discovery pipelines, combining genomics, transcriptomics, and metabolomics data with literature knowledge to identify novel biomarkers, drug targets, and therapeutic mechanisms. These integrated approaches demonstrate particular promise in precision medicine applications where molecular data interpretation benefits from contextual literature knowledge, enabling personalized therapeutic strategies tailored to specific patient subgroups while enhancing efficacy and minimizing adverse effects **(Figure 4)** **(Wang *et al.* 2021)**.

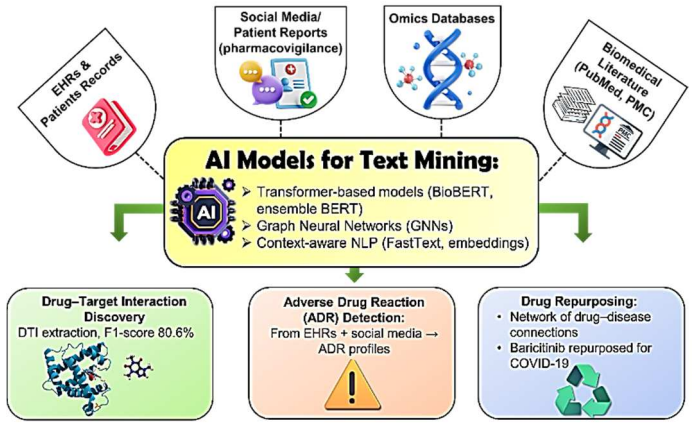

Figure 4: AI-Enabled Text Mining in Drug Discovery and Repurposing

7. Applications in Clinical Research and Decision Support

Automated systematic reviews and meta-analyses have achieved remarkable efficiency gains through AI implementation, fundamentally transforming evidence synthesis workflows. Automated meta-analysis (AMA) systems demonstrate 57% focus on data processing tasks including extraction and statistical modeling, while only 17% address advanced synthesis stages. AI-driven literature screening achieves combined recall of 92.8% and specificity of 64.7% when maximizing recall, though precision remains at 20.0%. Recent advances show AI tools can reduce literature screening time by up to 70% while maintaining acceptable accuracy. Sophisticated platforms like search refiner automatically identify frequent MeSH terms from selected references, categorizing them into health conditions, treatments, and study designs for Boolean query development. However, current evidence indicates manual literature screening remains indispensable for medical systematic reviews, with human oversight essential for ensuring methodological rigor **(Riaz *et al.* 2024)**.

Clinical trial conscription optimisation via NLP has confirmed substantial improvements in enrolment efficiency and candidate identification precision. AI-powered systems like IBM Watson Health's Clinical Trial Matching achieve 79.9% lessening in patients required for screening through computerized EHR text mining for cardiovascular trials. NLP algorithms managing electronic health records achieve 95% sensitivity

and 86% positive predictive value in finding heart failure patients with conserved ejection fraction for trial suitability. Innovative deep learning-based eligibility principles optimisation qualifies comprehensive protocol design enhancements while precisely recognizing suitable patients through integrated EHR analysis. Machine learning methods applying demographic data, presentation information, and admission actions accomplish performance close to complete EMR-based models, signifying effective enlistment through administrative data alone. Social media incorporation and digital campaigning through targeted platforms improve recruitment reach, with AI-driven patient matching systems improving enrolment rates across oncology and cardiovascular domains **(Alexander *et al.* 2020)**.

Enhanced Clinical Decision Support Systems (CDSS) through AI integration have achieved significant improvements in diagnostic accuracy and treatment recommendations. AI-powered CDSS utilizing machine learning algorithms demonstrate superior performance compared to traditional methods in disease diagnosis, leading to more accurate treatment suggestions and improved patient outcomes. Natural language processing engines effectively guide development of more accurate clinical decision systems, with biphasic tools combining AI algorithms and human criteria improving clinical diagnosis and treatment workflows. Deep learning models analyzing clinical narratives achieve enhanced diagnostic precision through sentiment analysis and entity recognition, identifying patient attitudes, preferences, and treatment adherence patterns. However, implementation challenges persist, with 79% of TM-CDS combination tools remaining unimplemented in clinical practice due to natural language complexity, EHR incompleteness, validation requirements, and adoption barriers **(Karuppan Perumal *et al.* 2025)**.

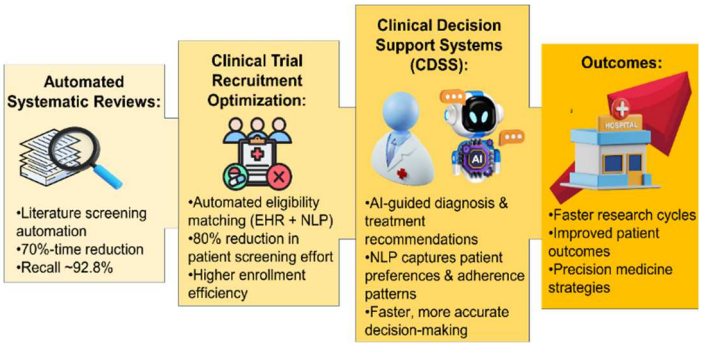

Figure 5: Applications of AI-Enabled Text Mining in Clinical Research and Decision Support

Patient stratification and personalized medicine approaches represent the most promising applications of AI-enabled text mining in precision healthcare. Machine learning techniques achieve 98.1% and 99.9% accuracy in COVID-19 patient stratification using integrated omics and clinical biomarker datasets, enabling precise prediction of disease severity and outcomes. AI algorithms significantly investigating the patient medical history, treatment response, lifestyle factors, and genetics that responsible for predicting the personalized treatment approach with reduced side effects via continuous learning from patients' outcomes. Deep learning approaches like NLP models achieve 92.6% sensitivity in complex clinical outcomes including goals-of-care discussions from HER text, qualifying large-scale pragmatic trials previously impracticable via manual abstraction **(Bello *et al.* 2023)**. Advanced patient risk stratification models have been applied to interpret the demographic data and clinician-initiated information to achieve performance close to complete EMR-based application, suggesting that effective personalization via accessible datasets. The incorporation of NLP with genomic data denotes the comprehensive information of patient phenotyping to identify treatment-responsive subgroups and for therapeutic optimization strategies across heterogeneous disease populations. These AI-based applications signify the precise efficacy in autoimmune diseases, chronic conditions, oncology where distinct pathogenetic signature are must for personalize the therapeutic inventions beyond traditional one size-fits-all approaches **(Figure 5) (Beaulieu-Jones *et al.* 2021)**.

8. Ethical, Legal, and Social Implications (ELSI)

Processing the medical data requires strict privacy and rules, HIPAA regulations it requires removal of 18 specific categories of Protected Health Information (PHI) including names, dates, and biometric identifiers. Whereas in GDPR extends beyond HIPAA's rules and focus to encompass all the details including personal data with explicit consent requirements for data usage and data breaches within 72hrs. these regulations ensures the personal data is safeguarded during processing the medical text mining **(Theodos and Sittig 2021)**. To handling the data privacy, the main challenges lie between the data utility and protecting patients' information and maintaining usefulness of the data. De-identification approaches including addition of statistical noise to datasets can reduce the analytical accuracy. Advanced privacy-preserving methods employed hierarchical access controls with five user levels from obfuscated data users receiving randomized aggregated data to PHI-viewable data users with complete access. Though, the deidentification approach can protect patients' privacy, still permits for re-identification when the datasets are combined, posing significant privacy risks. For example, China's personal information Protection Law (PIPLRC) allows automated decision making using anonymized health information with consumer consent demonstrating evolving global privacy frameworks **(Im *et al.* 2024)**. AI bias and fairness disparities signify critical threats to reasonable healthcare delivery, with biased algorithms thoroughly disadvantaging marginalized populations. For instance, Obermeyer's landmark study reported how healthcare resource allocation algorithms preferred recovering white patients over sicker Black patients owing to biased cost-based training data that disastrous to replicate true healthcare needs. Additionally, diagnostic AI systems show performance disparities across racial groups, with chest radiograph analysis algorithms showing reduced accuracy for minority populations. Biased AI system in healthcare leads the inefficient allocation and affects the medical staffing, diagnostic distribution and hospital bed assignments, which increase wait times and healthcare costs **(Ueda *et al.* 2024)**. Root causes of these issues include unrepresentative training datasets, development practices, and structural deficiencies within healthcare systems. Employing fairness-aware learning approaches signifies the measurable success with balanced datasets to achieve effective bias reduction via demographic parity an equalized odds metrics via complete elimination of bias remains challenging **(Chinta *et al.* 2025)**. Comprehensive approaches combination data preprocessing, algorithm modifications, and post-analysis intrusions create optimal fairness outcomes. The complexity of both explainability and interpretability poses significant challenges and barriers on clinical adoption of AI system, particularly in high-stakes decision making context. Medical AI applications require different stakeholders and varying level of explanation to ethics committees requires complex moral concept interpretations **(Amann *et al.* 2022)**. Post hoc explanations tools like LIME and SHAP offers useful insights in clinical settings to achieve restrained clinical utility, while knowledge-based hybrid models offer higher interpretability by integrating reputable medical knowledge frameworks. The accuracy-explainability trade off differs remarkably across use cases, emergency call systems arrange speed over interpretability, while psychiatric diagnosis tools need high explainability for junior clinicians **(Ennab and McHeick 2024)**. Automation bias risks distress all AI applications, with clinicians possibly over-trusting on AI recommendations without critical evaluation. Research reveals scarcity of empirical data on explainability effectiveness,

with most evidence remaining theoretical rather than demonstrating real-world clinical impact. Implementation challenges include natural language complexity, EHR data incompleteness, and validation requirements, resulting in 79% of explainable AI tools remaining undeployed in clinical practice **(Goddard *et al.* 2012)**. Advanced XAI approaches utilizing pre-modeling explainability through data visualization, inherently interpretable model architectures, and post-hoc explanation techniques show promise for enhancing clinical trust and decision-making accuracy. However, successful deployment requires use-case-specific studies investigating both technical explainability capabilities and clinician/patient perspectives on explanation relevance and utility.

9. Current Challenges and Limitations

Data heterogeneity and interoperability issues represent fundamental barriers to effective medical text mining deployment. Healthcare systems demonstrate persistent lack of standardization with multiple data formats and coding systems creating non-interoperable silos. EHR systems utilize proprietary formats that alter incoming data by mapping to internal, non-standard terms, creating complex data conversion cycles that degrade analytical quality. Even with standards like HL7 FHIR, defective adoption and varied application delay seamless data exchange, different retailers support various FHIR versions or interpretation diverse resources, meaning "FHIR-compliant" doesn't guarantee interoperability. More challenges are semantic discrepancies, where medical system often use different formats for same medical concepts across the system, which requires the extensive mapping and normalization. The surge of data from devices like phones and wearables and genetic testing makes this even harder because EMR systems weren't designed to incorporate diverse data sources. These heterogeneity issues significantly influence the AI effectiveness and poor data quality degrades machine learning model performance and inconsistent data producing unreliable results **(Torab-Miandoab *et al.* 2023)**. High-quality labels dataset plays a crucial role in in successful establishment of medical NLP, but annotation quality often varies widely. The inconsistency and poor lables can cost healthcare organizations 15% of revenue. It also faces specific challenges stemming from specialized terminology, poor documentation practices and the requirement for domain expertise. Manual annotations remain time-consuming and cost-effective method, while automated applications sacrifice precision, produces tension between accuracy and scalability. To practice consistent annotation for the medical annotation is very challenging because different people often label the same annotation differently and require extensive training and standardized protocols. There are very few gold-standard datasets for these issues and several studies use single physician's annotations as reference which can introducing potential bias. Domain-specific annotations requirements vary remarkably across medical specialties, necessitating specific training for every clinical framework. The recent advances in large language model-based annotations prediction exhibits the promise for generating custom domain-aligned datasets, still, validation remains challenging **(Sylolypavan *et al.* 2023)**. Scaling advanced text-mining system demands significant implementation barriers for healthcare organizations those have limited IT support. For instance, MERLIN a platform that process 15,000 patients records per hour and each has thousands of mixed data types relies on a microservice architecture. But in tests, it required 8-core CPUs, 6GB RAM per node for successful run. The volume of data and number of features extraction increases the processing complexity, for instance, the processing of 1000 computed variables required approximately 2.5 minutes, whereas the manual approach can take days to weeks. Real-time processing presents the additional issues, as tools designed for immediate clinical decision supports computational efficiency with analytical depth **(Cohen and Kovacheva 2023)**. Infrastructure costs signify substantial barriers for academic and smaller healthcare institutions; however, cloud-based solutions offer scalability without the need for substantial upfront hardware investments. Optimizing memory and CPU usage is critical when scaling, stateless microservices architecture are effective because it allows CPU usage to enhance the linearly with demand, however, the translating research finding in clinical practice demonstrates the most persistent challenge in medical AI deployment. Despite demonstrating efficacy in research settings, about 79% of the text mining clinical decision support tools have not been implemented in the real-world clinical practice owing to several barriers. Single-center studies have limited diversity and making it hard to apply the methodological framework broadly. When AI models trained in controlled settings, were applied in real-world clinical environments, their performance often declines owing to difference in patient populations and clinical practices**(Sachdeva *et al.* 2024)**. In addition, the incorporation of AI tools into existing workflows is challenging. Tools which required remarkable changes to development or have unfamiliar interfaces or disrupting established clinical processes face low adaptation in research settings. Implementing AI tools in healthcare needs substantial and change management efforts, which often exceed the organization's capacity, with lack of AI literacy among clinicians representing a fundamental implementation barrier. Financial sustainability remains unclear for many AI applications, with uncertain return on investment limiting institutional commitment. Success involves complicated methods combining technical validation, workflow optimization, regulatory clarity, and complete training programs to bridge the persistent research-to-practice divide **(Hirani *et al.* 2024)**.

10. Future Directions and Emerging Trends

Multimodal text mining combination indicates the next frontier in precision medicine, demonstrating outstanding latent through combination of diverse biomedical data streams. AI-enabled advanced multimodal achieve 86.05% accuracy in schizophrenia classification by incorporating the genomic, structural MRI, and functional connectivity via transformer structures. In oncology applications, this multimodal system achieves are under curve of 0.91 for predicting anti-HER2 therapeutic approaches by combining both pathological imaging and omics data. For instance, the MIGTrans architecture employed the genomic encodes, connectome analysis and spatial sequence attention to analysis the comprehensive neuroanatomical abnormalities and outperforming single modality approaches **(Huang and Shu 2025)**. In case of cardiovascular risk assessment, to achieve the highest performance over traditional Framingham Risk Score it incorporates the following parameters including echocardiograms, genetic data, electronic health records. Hence, the future development highly focused on integrating the wearable biosensor data and environmental factors while addressing the interoperability challenges and achieve consistent algorithmic transparency **(Singh *et al.* 2024)**. Multilingual corpora, regional benchmark datasets, and cross-lingual transfer learning are utilized to validate text-mining models across various languages, while ontologies such as UMLS, MeSH, and GO standardize multi-omics integration to guarantee consistent assessment across a wide range of clinical and molecular datasets.

Federated learning and privacy-preserving AI have emerged as critical enablers for large-scale healthcare collaboration without compromising data security. FL systems demonstrate superior model generalizability across diverse healthcare institutions while maintaining local data privacy through encrypted model update exchanges. Privacy-preserving techniques including homomorphic encryption and secure multi-party computation enable collaborative AI training with minimal information leakage risk. Medical imaging applications show particular promise, with federated approaches achieving comparable performance to centralized

models while ensuring HIPAA and GDPR compliance. However, challenges persist including trust establishment among participating institutions, model update information leakage, and computational overhead from encryption protocols. Advanced differential privacy mechanisms and hierarchical access control systems provide additional security layers, though implementation complexity remains a barrier to widespread adoption **(Loftus *et al.* 2022)**. Real-time clinical decision-making with large language models has achieved remarkable efficiency gains across healthcare workflows. LLM systems reduce clinical decision time from 33.60 minutes for expert multidisciplinary teams to under 1 minute for top-performing models like GPTo1, GPT4o, and Deepseek, while maintaining comparable diagnostic accuracy with mean Likert scores exceeding 4.19. Advanced retrieval-augmented generation (RAG) pipelines improve question-answering accuracy by up to 16.6% while significantly reducing hallucinations through structured medical knowledge integration **(Gaber *et al.* 2025)**. Five-stage clinical workflow integration demonstrates LLM utility across patient registration, examination coordination, diagnosis formulation, treatment administration, and discharge planning. Real-time triage optimization shows potential for reducing both under-triage (14-34%) and over-triage (12-31%) rates, improving resource allocation and patient outcomes. However, infrastructure requirements, data privacy compliance, and interoperability with existing health information systems remain significant implementation challenges **(Li *et al.* 2025)**. Generative AI in hypothesis generation and drug innovation represents a transformative paradigm shift from traditional hypothesis-driven approaches. FRONTEO's Drug Discovery AI Factory analyzes over 30 million PubMed reports to identify highly novel target molecules previously undetected by conventional gene expression analysis and GWAS methods. AI-driven approaches enable parallel analysis of multiple development stages, automated data processing, and multimodal data utilization compared to linear, sequential traditional methods. Hypothesis generation encompasses target molecule identification, disease mechanism elucidation, patient stratification, safety profiling, and experimental model proposals within integrated platforms. Drug repositioning applications demonstrate significant acceleration in identifying novel therapeutic applications through comprehensive literature mining and network analysis. ChatGPT-based systems show promising capabilities in generating innovative research hypotheses for addressing complex challenges like cardiotoxicity, though validation through human expertise remains essential. The integration of hypothesis generation with predictive analytics enables early anticipation of drug success rates and potential adverse effects, potentially reducing high attrition rates characteristic of traditional development **(Gangwal and Lavecchia 2024)**. Explainable and trustworthy AI in medical text mining has become increasingly critical for clinical adoption and regulatory compliance. Advanced XAI approaches encompass three main strategies: pre-modeling explainability through data visualization, inherently interpretable model architectures, and post-hoc explanation techniques. Medical applications require multiple explainability levels addressing different stakeholders, from emergency dispatchers needing minimal explanations to ethics committees requiring complex moral concept interpretations. LIME and SHAP-based post-hoc explanations achieve moderate clinical utility, while knowledge-based hybrid models provide superior interpretability through established medical knowledge frameworks **(Di Martino and Delmastro 2023)**. However, accuracy-explainability tradeoffs vary significantly across use cases, with emergency systems prioritizing speed over interpretability while psychiatric diagnosis tools require high explainability for junior clinicians. Implementation challenges include 79% of explainable AI tools remaining undeployed due to natural language complexity, validation requirements, and adoption barriers. Future developments focus on use-case-specific explainability studies investigating both technical capabilities and clinician/patient perspectives on explanation relevance, utility, and trust-building potential **(Okada *et al.* 2023)**.

11. Conclusion

AI-enabled text mining has rapidly evolved from a promising research tool to a clinically relevant force driving transformation across healthcare. By unlocking the vast reservoir of unstructured biomedical data, clinical notes, research literature, trial records, and patient communications, it bridges the gap between data abundance and human interpretive capacity. This integration of artificial intelligence with natural language processing enables actionable insights that are reshaping disease prediction, accelerating drug discovery, and streamlining clinical research. A number of studies have been reported that the perceptible clinical values of the AI approach for the early diagnosis of diseases. AI-applications consistently outperform conventional scoring systems and drug-target binding mechanisms achieves robust accuracy and automated evidence synthesis reduces the systematic review timelines by over half while maintaining consistency. In corresponding, real-time decision supported by large language models transports near-expert diagnostic presentation in seconds, while the multimodal incorporation of text with sensor data, imaging and genomics significantly improved the personalized care strategies. Also, AI-based text mining significantly involved in democratizes biomedical intelligence for researchers to frame the large-scale hypothesis and collaboration with institution via privacy-preserving frameworks including federated learning. Hence, its efficacy extends into public health, where the outbreak detection and real-world evidence mining provides proactive strategies for policy intervention and surveillance. Though, the persistent challenges in algorithmic bias, explainability and workflow integration and regulatory clarity should be addressed to certify the equitable and reliable adoption. Incorporating the explainable AI, multimodal data mining and system which propose new strategies promise a paradigm shift making it more predictive, preventive and personalized medicine. Prominently, the true measure of success will not lie in computational complexity but in noticeable enhancements in patient outcomes, health equity, and system efficiency. As healthcare enters this new digital frontier, the critical task gaining is positioning infrastructure, policies, and workforce training to participate AI with human proficiency, certifying that technological progress improves, rather than substitutes, compassionate medical care.

12. Disclosure Statements

12.1. Author Contribution

YS: Data collection, manuscript drafting, Data curation, literature review, manuscript preparation. **JM:** Conceptualization, study design, supervision, data interpretation, manuscript writing and revision. The corresponding author have read and approved the final manuscript.

12.2. Declaration of Generative AI

The authors declare that no generative AI tools were used in the drafting, writing, or editing of the manuscript. All scientific interpretations and conclusions are the author's own.

12.3. Ethics approval (for clinical/animal studies)

This article is a review paper and does not involve any original research on human participants, clinical samples, or experimental animals. All data and information presented are derived from previously published studies. Therefore, ethical approval from an institutional review board or ethics committee is not required.

12.4. Informed Consent Statement

Not applicable.

12.5. Data Availability Statement

The study didn't proceed with any data generation.

12.6. Acknowledgment

The authors thankfully acknowledge the Electric Power Research Institute, 1300 W W T Harris Blvd, Charlotte, NC, USA for providing necessary facilities for performing this study.

12.7. Funding Statement

This research received no external funding. The study was conducted without any financial support from public, commercial, or not-for-profit funding agencies. All resources utilized for this work were provided by the author respective institutions.

12.8. Conflicts of Interest

The authors declare that they have no known financial, personal, academic, or other relationships that could inappropriately influence, or be perceived to influence, the work reported in this manuscript. All authors confirm that there are no competing interests to declare.

12.9. Corresponding Author Contact Information

The corresponding author **Dr. M. Jayakumar** can be contacted via email jmanoharan[at]outlook.com.

12.10. Supplementary Information

No supplementary material is available for this article.

12.11. ORcID Information

Manoharan 0009-0009-7765-9165
Sehgal 0009-0004-2364-2310

12.12. Handling Editor Information

This manuscript was handled and edited by **Dr. Chandrabose Selvaraj**, Professor, Bioinformatics Division, Department of Marine Biotechnology, AMET University (Academy of Maritime Education and Training), Deemed to be University, East Coast Road, Kanathur, Chennai, Tamil Nadu – 603112, India. **Editor contact email**: jomi[at]aayvu.com

13. Reference

Ahmad PN, Shah AM, Lee K. (2023), A Review on Electronic Health Record Text-Mining for Biomedical Name Entity Recognition in Healthcare Domain, *Healthcare (Basel),* 11(9) **doi:**10.3390/healthcare11091268. **PMID:** 37174810.

Al-Abri R, Al-Balushi A. (2014), Patient satisfaction survey as a tool towards quality improvement, *Oman Med J,* 29(1):3-7. **doi:**10.5001/omj.2014.02. **PMID:** 24501659.

Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, Subbarayalu AV, Al Muhanna D, Al-Muhanna FA. (2023), A Review of the Role of Artificial Intelligence in Healthcare, *J Pers Med,* 13(6) **doi:**10.3390/jpm13060951. **PMID:** 37373940.

Aldahdooh J, Tanoli Z, Tang J. (2024), Mining drug-target interactions from biomedical literature using chemical and gene descriptions-based ensemble transformer model, *Bioinform Adv,* 4(1):vbae106. **doi:**10.1093/bioadv/vbae106. **PMID:** 39092007.

Alexander M, Solomon B, Ball DL, Sheerin M, Dankwa-Mullan I, Preininger AM, Jackson GP, Herath DM. (2020), Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients, *JAMIA Open,* 3(2):209-215. **doi:**10.1093/jamiaopen/ooaa002. **PMID:** 32734161.

Algera MD, Morton R, Sundar SS, Farrell R, van Driel WJ, Brennan D, Rijken MJ, Sfeir S, Allen L, Eiken M, Coleman RL, collaborators of the Global Equality in Ovarian Cancer Care project g. (2023), Exploring international differences in ovarian cancer care: a survey report on global patterns of care, current practices, and barriers, *Int J Gynecol Cancer,* 33(10):1612-1620. **doi:**10.1136/ijgc-2023-004563. **PMID:** 37591611.

Almeman A. (2024), The digital transformation in pharmacy: embracing online platforms and the cosmeceutical paradigm shift, *J Health Popul Nutr,* 43(1):60. **doi:**10.1186/s41043-024-00550-2. **PMID:** 38720390.

Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, Gilbert TK, Hagendorff T, Holm S, Livne M, Spezzatti A, Strumke I, Zicari RV, Madai VI, initiative ZI. (2022), To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems, *PLOS Digit Health,* 1(2):e0000016. **doi:**10.1371/journal.pdig.0000016. **PMID:** 36812545.

Awrahman BJ, Aziz Fatah C, Hamaamin MY. (2022), A Review of the Role and Challenges of Big Data in Healthcare Informatics and Analytics, *Comput Intell Neurosci,* 2022:5317760. **doi:**10.1155/2022/5317760. **PMID:** 36210978.

Bajwa J, Munir U, Nori A, Williams B. (2021), Artificial intelligence in healthcare: transforming the practice of medicine, *Future Healthc J,* 8(2):e188-e194. **doi:**10.7861/fhj.2021-0095. **PMID:** 34286183.

Batko K, Slezak A. (2022), The use of Big Data Analytics in healthcare, *J Big Data,* 9(1):3. **doi:**10.1186/s40537-021-00553-4. **PMID:** 35013701.

Beaulieu-Jones BK, Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, Kohane IS. (2021), Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?, *NPJ Digit Med,* 4(1):62. **doi:**10.1038/s41746-021-00426-3. **PMID:** 33785839.

Bello B, Bundey YN, Bhave R, Khotimchenko M, Baran SW, Chakravarty K, Varshney J. (2023), Integrating AI/ML Models for Patient Stratification Leveraging Omics Dataset and Clinical Biomarkers from COVID-19 Patients: A Promising Approach to Personalized Medicine, *Int J Mol Sci,* 24(7) **doi:**10.3390/ijms24076250. **PMID:** 37047222.

Cheng Y, Cheng R, Xu T, Tan X, Bai Y. (2025), Machine Learning Techniques Applied to COVID-19 Prediction: A Systematic Literature Review, *Bioengineering (Basel),* 12(5) **doi:**10.3390/bioengineering12050514. **PMID:** 40428133.

Chinta SV, Wang Z, Palikhe A, Zhang X, Kashif A, Smith MA, Liu J, Zhang W. (2025), AI-driven healthcare: Fairness in AI healthcare: A survey, *PLOS Digit Health,* 4(5):e0000864. **doi:**10.1371/journal.pdig.0000864. **PMID:** 40392801.

Chong PL, Vaigeshwari V, Mohammed Reyasudin BK, Noor Hidayah BRA, Tatchanaamoorti P, Yeow JA, Kong FY. (2025), Integrating artificial intelligence in healthcare: applications, challenges, and future directions, *Future Sci OA,* 11(1):2527505. **doi:**10.1080/20565623.2025.2527505. **PMID:** 40616302.

Cohen RY, Kovacheva VP. (2023), A Methodology for a Scalable, Collaborative, and Resource-Efficient Platform, MERLIN, to Facilitate Healthcare AI Research, *IEEE J Biomed Health Inform,* 27(6):3014-3025. **doi:**10.1109/JBHI.2023.3259395. **PMID:** 37030761.

Dai L, Xu H, Zhang Y. (2025), Automated classification of clinical diagnoses in electronic health records using transformer, *PLoS One,* 20(9):e0329963. **doi:**10.1371/journal.pone.0329963. **PMID:** 40934268.

Denecke K, Reichenpfader D. (2023), Sentiment analysis of clinical narratives: A scoping review, *J Biomed Inform,* 140:104336. **doi:**10.1016/j.jbi.2023.104336. **PMID:** 36958461.

Dermawan D, Alotaiq N. (2025), From Lab to Clinic: How Artificial Intelligence (AI) Is Reshaping Drug Discovery Timelines and

Industry Outcomes, *Pharmaceuticals (Basel),* 18(7) **doi:**10.3390/ph18070981. **PMID:** 40732273.

Di Martino F, Delmastro F. (2023), Explainable AI for clinical and remote health applications: a survey on tabular and time series data, *Artif Intell Rev,* 56(6):5261-5315. **doi:**10.1007/s10462-022-10304-3. **PMID:** 36320613.

El-Warrak L, Nunes M, Luna G, Barbosa CE, Lyra A, Argolo M, Lima Y, Salazar H, de Souza JM. (2023), Towards the Future of Public Health: Roadmapping Trends and Scenarios in the Post-COVID Healthcare Era, *Healthcare (Basel),* 11(24) **doi:**10.3390/healthcare11243118. **PMID:** 38132008.

Ennab M, McHeick H. (2024), Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions, *Front Robot AI,* 11:1444763. **doi:**10.3389/frobt.2024.1444763. **PMID:** 39677978.

Friedman JI, Parchure P, Cheng FY, Fu W, Cheertirala S, Timsina P, Raut G, Reina K, Joseph-Jimerson J, Mazumdar M, Freeman R, Reich DL, Kia A. (2025), Machine Learning Multimodal Model for Delirium Risk Stratification, *JAMA Netw Open,* 8(5):e258874. **doi:**10.1001/jamanetworkopen.2025.8874. **PMID:** 40332938.

Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, Franke V, Akalin A. (2025), Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis, *NPJ Digit Med,* 8(1):263. **doi:**10.1038/s41746-025-01684-1. **PMID:** 40346344.

Gangwal A, Lavecchia A. (2024), Unleashing the power of generative AI in drug discovery, *Drug Discov Today,* 29(6):103992. **doi:**10.1016/j.drudis.2024.103992. **PMID:** 38663579.

Ge L, Agrawal R, Singer M, Kannapiran P, De Castro Molina JA, Teow KL, Yap CW, Abisheganaden JA. (2024), Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges, *Syst Rev,* 13(1):269. **doi:**10.1186/s13643-024-02682-2. **PMID:** 39456077.

Goddard K, Roudsari A, Wyatt JC. (2012), Automation bias: a systematic review of frequency, effect mediators, and mitigators, *J Am Med Inform Assoc,* 19(1):121-7. **doi:**10.1136/amiajnl-2011-000089. **PMID:** 21685142.

Gonzalez G, Cohen KB, Greene CS, Hahn U, Kann MG, Leaman R, Shah N, Ye J. (2013), Text and data mining for biomedical discovery, *Pac Symp Biocomput,* 2013:368-72. **doi:**10.1142/9789814583220_0030. **PMID:** 23424141.

Hirani R, Noruzi K, Khuram H, Hussaini AS, Aifuwa EI, Ely KE, Lewis JM, Gabr AE, Smiley A, Tiwari RK, Etienne M. (2024), Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities, *Life (Basel),* 14(5) **doi:**10.3390/life14050557. **PMID:** 38792579.

Hossain S, Hasan MK, Faruk MO, Aktar N, Hossain R, Hossain K. (2024), Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023, *BMC Cardiovasc Disord,* 24(1):214. **doi:**10.1186/s12872-024-03883-2. **PMID:** 38632519.

Huang W, Shu N. (2025), AI-powered integration of multimodal imaging in precision medicine for neuropsychiatric disorders, *Cell Rep Med,* 6(5):102132. **doi:**10.1016/j.xcrm.2025.102132. **PMID:** 40398391.

Im E, Kim H, Lee H, Jiang X, Kim JH. (2024), Exploring the tradeoff between data privacy and utility with a clinical data analysis use case, *BMC Med Inform Decis Mak,* 24(1):147. **doi:**10.1186/s12911-024-02545-9. **PMID:** 38816848.

Jadczyk T, Wojakowski W, Tendera M, Henry TD, Egnaczyk G, Shreenivas S. (2021), Artificial Intelligence Can Improve Patient Management at the Time of a Pandemic: The Role of Voice Technology, *J Med Internet Res,* 23(5):e22959. **doi:**10.2196/22959. **PMID:** 33999834.

Jin S, Niu Z, Jiang C, Huang W, Xia F, Jin X, Liu X, Zeng X. (2021), HeTDR: Drug repositioning based on heterogeneous networks and text mining, *Patterns (N Y),* 2(8):100307. **doi:**10.1016/j.patter.2021.100307. **PMID:** 34430926.

Kanchan S, Gaidhane A. (2023), Social Media Role and Its Impact on Public Health: A Narrative Review, *Cureus,* 15(1):e33737. **doi:**10.7759/cureus.33737. **PMID:** 36793805.

Karuppan Perumal MK, Rajan Renuka R, Kumar Subbiah S, Manickam Natarajan P. (2025), Artificial intelligence-driven clinical decision support systems for early detection and precision therapy in oral cancer: a mini review, *Front Oral Health,* 6:1592428. **doi:**10.3389/froh.2025.1592428. **PMID:** 40356851.

Kumar R, Garg S, Kaur R, Johar MGM, Singh S, Menon SV, Kumar P, Hadi AM, Hasson SA, Lozanovic J. (2025), A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions, *Front Artif Intell,* 8:1583459. **doi:**10.3389/frai.2025.1583459. **PMID:** 40433606.

Lee C, Britto S, Diwan K. (2024), Evaluating the Impact of Artificial Intelligence (AI) on Clinical Documentation Efficiency and Accuracy Across Clinical Settings: A Scoping Review, *Cureus,* 16(11):e73994. **doi:**10.7759/cureus.73994. **PMID:** 39703286.

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. (2020), BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics,* 36(4):1234-1240. **doi:**10.1093/bioinformatics/btz682. **PMID:** 31501885.

Li H, Fu JF, Python A. (2025), Implementing Large Language Models in Health Care: Clinician-Focused Review With Interactive Guideline, *J Med Internet Res,* 27:e71916. **doi:**10.2196/71916. **PMID:** 40644686.

Loftus TJ, Ruppert MM, Shickel B, Ozrazgat-Baslanti T, Balch JA, Efron PA, Upchurch GR, Jr., Rashidi P, Tignanelli C, Bian J, Bihorac A. (2022), Federated learning for preserving data privacy in collaborative healthcare research, *Digit Health,* 8:20552076221134455. **doi:**10.1177/20552076221134455. **PMID:** 36325438.

Maleki Varnosfaderani S, Forouzanfar M. (2024), The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century, *Bioengineering (Basel),* 11(4) **doi:**10.3390/bioengineering11040337. **PMID:** 38671759.

Modi S, Feldman SS. (2022), The Value of Electronic Health Records Since the Health Information Technology for Economic and Clinical Health Act: Systematic Review, *JMIR Med Inform,* 10(9):e37283. **doi:**10.2196/37283. **PMID:** 36166286.

Moon S, Pakhomov S, Melton GB. (2012), Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations, *AMIA Annu Symp Proc,* 2012:1310-9. **doi:**https://www.ncbi.nlm.nih.gov/pubmed/23304410. **PMID:** 23304410.

Murray C, Mitchell L, Tuke J, Mackay M. (2024), Revealing patient-reported experiences in healthcare from social media using thedesign-acquire-process-model-analyse-visualise framework, *Digit Health,* 10:20552076241251715. **doi:**10.1177/20552076241251715. **PMID:** 38757085.

Okada Y, Ning Y, Ong MEH. (2023), Explainable artificial intelligence in emergency medicine: an overview, *Clin Exp Emerg Med,* 10(4):354-362. **doi:**10.15441/ceem.23.145. **PMID:** 38012816.

Riaz IB, Naqvi SAA, Hasan B, Murad MH. (2024), Future of Evidence Synthesis: Automated, Living, and Interactive Systematic Reviews and Meta-analyses, *Mayo Clin Proc Digit Health,* 2(3):361-365. **doi:**10.1016/j.mcpdig.2024.05.023. **PMID:** 40206128.

Rocha HAL, Solha EZM, Furtado V, Justino FL, Barreto LAL, da Silva RG, de Oliveira IM, Bates DW, de Goes Cavalcanti LP, Lima Neto AS, de

Oliveira EA. (2024), COVID-19 outbreaks surveillance through text mining applied to electronic health records, *BMC Infect Dis,* 24(1):359. **doi:**10.1186/s12879-024-09250-y. **PMID:** 38549109.

Sachdeva S, Bhatia S, Al Harrasi A, Shah YA, Anwer K, Philip AK, Shah SFA, Khan A, Ahsan Halim S. (2024), Unraveling the role of cloud computing in health care system and biomedical sciences, *Heliyon,* 10(7):e29044. **doi:**10.1016/j.heliyon.2024.e29044. **PMID:** 38601602.

Schinkel M, van der Poll T, Wiersinga WJ. (2023), Artificial Intelligence for Early Sepsis Detection: A Word of Caution, *Am J Respir Crit Care Med,* 207(7):853-854. **doi:**10.1164/rccm.202212-2284VP. **PMID:** 36724366.

Seyedtabib M, Najafi-Vosough R, Kamyari N. (2024), The predictive power of data: machine learning analysis for Covid-19 mortality based on personal, clinical, preclinical, and laboratory variables in a case-control study, *BMC Infect Dis,* 24(1):411. **doi:**10.1186/s12879-024-09298-w. **PMID:** 38637727.

Shameer K, Johnson KW, Yahi A, Miotto R, Li LI, Ricks D, Jebakaran J, Kovatch P, Sengupta PP, Gelijns S, Moskovitz A, Darrow B, David DL, Kasarskis A, Tatonetti NP, Pinney S, Dudley JT. (2017), Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-Wide Machine Learning: A Case-Study Using Mount Sinai Heart Failure Cohort, *Pac Symp Biocomput,* 22:276-287. **doi:**10.1142/9789813207813_0027. **PMID:** 27896982.

Singh M, Kumar A, Khanna NN, Laird JR, Nicolaides A, Faa G, Johri AM, Mantella LE, Fernandes JFE, Teji JS, Singh N, Fouda MM, Singh R, Sharma A, Kitas G, Rathore V, Singh IM, Tadepalli K, Al-Maini M, Isenovic ER, Chaturvedi S, Garg D, Paraskevas KI, Mikhailidis DP, Viswanathan V, Kalra MK, Ruzsa Z, Saba L, Laine AF, Bhatt DL, Suri JS. (2024), Artificial intelligence for cardiovascular disease risk assessment in personalised framework: a scoping review, *EClinicalMedicine,* 73:102660. **doi:**10.1016/j.eclinm.2024.102660. **PMID:** 38846068.

Sylolypavan A, Sleeman D, Wu H, Sim M. (2023), The impact of inconsistent human annotations on AI driven clinical decision making, *NPJ Digit Med,* 6(1):26. **doi:**10.1038/s41746-023-00773-3. **PMID:** 36810915.

Tan JK, Zhang X, Cheng D, Leong IYO, Wong CS, Tey J, Loh SC, Soh EF, Lim WY. (2023), Using the Johns Hopkins ACG Case-Mix System for population segmentation in a hospital-based adult patient population in Singapore, *BMJ Open,* 13(3):e062786. **doi:**10.1136/bmjopen-2022-062786. **PMID:** 36997258.

Theodos K, Sittig S. (2021), Health Information Privacy Laws in the Digital Age: HIPAA Doesn't Apply, *Perspect Health Inf Manag,* 18(Winter):1l. **doi:**https://www.ncbi.nlm.nih.gov/pubmed/33633522. **PMID:** 33633522.

Theodosiou T, Vrettos K, Baltsavia I, Baltoumas F, Papanikolaou N, Antonakis A, Mossialos D, Ouzounis CA, Promponas VJ, Karaglani M, Chatzaki E, Brandau S, Pavlopoulos GA, Andreakos E, Iliopoulos I. (2024), BioTextQuest v2.0: An evolved tool for biomedical literature mining and concept discovery, *Comput Struct Biotechnol J,* 23:3247-3253. **doi:**10.1016/j.csbj.2024.08.016. **PMID:** 39279874.

Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. (2023), Interoperability of heterogeneous health information systems: a systematic literature review, *BMC Med Inform Decis Mak,* 23(1):18. **doi:**10.1186/s12911-023-02115-5. **PMID:** 36694161.

Tse T, Williams RJ, Zarin DA. (2009), Reporting "basic results" in ClinicalTrials.gov, *Chest,* 136(1):295-303. **doi:**10.1378/chest.08-3022. **PMID:** 19584212.

Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, Matsui Y, Nozaki T, Nakaura T, Fujima N, Tatsugami F, Yanagawa M, Hirata K, Yamada A, Tsuboyama T, Kawamura M, Fujioka T, Naganawa S. (2024), Fairness of artificial intelligence in healthcare: review and recommendations, *Jpn J Radiol,* 42(1):3-15. **doi:**10.1007/s11604-023-01474-3. **PMID:** 37540463.

Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S. (2019), Applications of machine learning in drug discovery and development, *Nat Rev Drug Discov,* 18(6):463-477. **doi:**10.1038/s41573-019-0024-5. **PMID:** 30976107.

van de Burgt BWM, Wasylewicz ATM, Dullemond B, Jessurun NT, Grouls RJE, Bouwman RA, Korsten EHM, Egberts TCG. (2024), Development of a text mining algorithm for identifying adverse drug reactions in electronic health records, *JAMIA Open,* 7(3):ooae070. **doi:**10.1093/jamiaopen/ooae070. **PMID:** 39156048.

van Mossel S, Oude-Wolcherink MJ, de Feria Cardet RE, de Geus-Oei LF, Vriens D, Koffijberg H, Saing S. (2025), Artificial Intelligence as a New Research Ally? Performing AI-Assisted Systematic Literature Reviews in Health Economics, *Pharmacoeconomics,* 43(6):647-650. **doi:**10.1007/s40273-025-01481-4. **PMID:** 40208557.

Wandy J, Daly R. (2021), GraphOmics: an interactive platform to explore and integrate multi-omics data, *BMC Bioinformatics,* 22(1):603. **doi:**10.1186/s12859-021-04500-1. **PMID:** 34922446.

Wang J, Wu Z, Peng Y, Li W, Liu G, Tang Y. (2021), Pathway-Based Drug Repurposing with DPNetinfer: A Method to Predict Drug-Pathway Associations via Network-Based Approaches, *J Chem Inf Model,* 61(5):2475-2485. **doi:**10.1021/acs.jcim.1c00009. **PMID:** 33900090.

Wang YJ, Choo WC, Ng KY, Bi R, Wang PW. (2025), Evolution of AI enabled healthcare systems using textual data with a pretrained BERT deep learning model, *Sci Rep,* 15(1):7540. **doi:**10.1038/s41598-025-91622-8. **PMID:** 40038367.

Xue C, Kowshik SS, Lteif D, Puducheri S, Jasodanand VH, Zhou OT, Walia AS, Guney OB, Zhang JD, Poesy S, Kaliaev A, Andreu-Arasa VC, Dwyer BC, Farris CW, Hao H, Kedar S, Mian AZ, Murman DL, O'Shea SA, Paul AB, Rohatgi S, Saint-Hilaire MH, Sartor EA, Setty BN, Small JE, Swaminathan A, Taraschenko O, Yuan J, Zhou Y, Zhu S, Karjadi C, Alvin Ang TF, Bargal SA, Plummer BA, Poston KL, Ahangaran M, Au R, Kolachalama VB. (2024), AI-based differential diagnosis of dementia etiologies on multimodal data, *Nat Med,* 30(10):2977-2989. **doi:**10.1038/s41591-024-03118-z. **PMID:** 38965435.

Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. (2023), Large language models in health care: Development, applications, and challenges, *Health Care Sci,* 2(4):255-263. **doi:**10.1002/hcs2.61. **PMID:** 38939520.

Yang Z, Kotoge R, Piao X, Chen Z, Zhu L, Gao P, Matsubara Y, Sakurai Y, Sun J. (2025), MLOmics: Cancer Multi-Omics Database for Machine Learning, *Sci Data,* 12(1):913. **doi:**10.1038/s41597-025-05235-x. **PMID:** 40447627.

You JG, Dbouk RH, Landman A, Ting DY, Dutta S, Wang JC, Centi AJ, Macfarlane M, Bechor E, Letourneau J, Choo-Kang G, Kim EH, Magee C, Lang BJ, Angelo L, Olin J, Frits M, Iannaccone C, Rui A, Salmikova I, Holland C, Blanchette B, Silverman R, Bates DW, Rotenstein L, Mishuris RG. (2025), Ambient Documentation Technology in Clinician Experience of Documentation Burden and Burnout, *JAMA Netw Open,* 8(8):e2528056. **doi:**10.1001/jamanetworkopen.2025.28056. **PMID:** 40839265.

Zakkar MA, Lizotte DJ. (2021), Analyzing Patient Stories on Social Media Using Text Analytics, *J Healthc Inform Res,* 5(4):382-400. **doi:**10.1007/s41666-021-00097-5. **PMID:** 35419510.

Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. (2011), The ClinicalTrials.gov results database--update and key issues, *N Engl J Med,* 364(9):852-60. **doi:**10.1056/NEJMsa1012065. **PMID:** 21366476.

Zhang YP, Zhang XY, Cheng YT, Li B, Teng XZ, Zhang J, Lam S, Zhou T, Ma ZR, Sheng JB, Tam VCW, Lee SWY, Ge H, Cai J. (2023), Artificial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling, *Mil Med Res,* 10(1):22. **doi:**10.1186/s40779-023-00458-8. **PMID:** 37189155.

Zhou N, Zhang CT, Lv HY, Hao CX, Li TJ, Zhu JJ, Zhu H, Jiang M, Liu KW, Hou HL, Liu D, Li AQ, Zhang GQ, Tian ZB, Zhang XC. (2019), Concordance Study Between IBM Watson for Oncology and Clinical Practice for Patients with Cancer in China, *Oncologist,* 24(6):812-819. **doi:**10.1634/theoncologist.2018-0255. **PMID:** 30181315.

How to Cite: Manoharan, J., and Sehgal, Y. (2026). AI-Enabled Text Mining: A Paradigm Shift in Disease Prediction, Drug Discovery, and Clinical Research. *Journal of Medico Informatics*, 02(02), 23-35. **doi:** https://doi.org/10.64659/jomi/215260