AI and Machine Learning in Clinical Medicine
Bridging or Separating Model Intelligence and Human Expertise

Fateme Nateghi Haredasht

Stanford Center for Biomedical Informatics Research,

Stanford University, Stanford, CA, USA

Email: fnateghi@stanford.edu

Joseph D. Romano

Department of Biostatistics, Epidemiology and Informatics,

University of Pennsylvania, Philadelphia, PA, USA

Email: joseph.romano@pennmedicine.upenn.edu

Brett K Beaulieu-Jones

Section of Computational Biomedicine and Biomedical Data Science,

University of Chicago, Chicago, IL, USA

Email: beaulieujones@uchicago.edu

Dokyoon Kim

Department of Biostatistics, Epidemiology and Informatics,

University of Pennsylvania, Philadelphia, PA, USA

Email:dokyoon.kim@pennmedicine.upenn.edu

Alexander Ioannidis

Department of Biomedical Data Science,

Stanford University, Stanford, CA, USA

Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA

Email: ioannid@stanford.edu

Geoffrey H Tison

Division of Cardiology, Center for Biosignal Research,

University of California, San Francisco, San Francisco, CA, USA

Email: geoff.tison@ucsf.edu

Roxana Daneshjou

Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

Email: roxanad@stanford.edu

Jonathan H. Chen

Department of Medicine and Center for Biomedical Informatics Research,

Stanford University, Stanford, CA, USA

Email: jonc101@stanford.edu

Artificial Intelligence (AI) technologies continue to expand their role in clinical medicine, with large language models (LLMs) and multimodal systems now applied to communication, imaging, and predictive analytics. Advances in generative and retrieval-augmented methods have improved the accuracy and contextual grounding of clinical summaries, patient messaging, and decision support. At the same time, new benchmarks in imaging, vision, and spontaneous speech have underscored both progress and the persistence of unsolved challenges. Predictive modeling efforts highlight causality, longitudinal trajectories, and informative clinical events, while methodological contributions emphasize uncertainty management, abstention, and interpretable causal structures. Finally, frameworks for evaluation and governance address the crucial gap between laboratory performance and real-world deployment.

Keywords: Artificial Intelligence, clinical medicine, decision support systems, large language models

1. Introduction

Healthcare stands at a critical point where computational approaches to medical decision-making are becoming increasingly sophisticated and integrated into clinical practice. The past decade has witnessed notable advancements in artificial intelligence (AI) and machine learning methods applied to healthcare challenges (Davenport and Kalakota, 2019; Rajkomar, Dean and Kohane, 2019). These innovations span numerous domains, including medical imaging interpretation, clinical decision support, predictive analytics for patient outcomes, and natural language processing (NLP) of electronic health records (EHRs) (Noei *et al.*, 2016; Ashtari, Maes and Van Huffel, 2021; Ashtari *et al.*, 2022; Nateghi Haredasht and Vens, 2022; Ashtari *et al.*, 2023; Nateghi Haredasht *et al.*, 2023; Nateghi Haredasht, Fouladvand, *et al.*, 2024; Bedi *et al.*, 2025; Chang *et al.*, 2025; Lopez *et al.*, 2025).

The rapid emergence of large language models (LLMs) and foundation models in healthcare has particularly accelerated in recent years, promising to transform how clinicians interact with medical information and make decisions (Bommasani *et al.*, 2022; Thirunavukarasu *et al.*, 2023). Studies have demonstrated their potential for clinical documentation, summarization of medical literature, patient triaging, and even direct patient interaction (Patel *et al.*, 2019; Nori *et al.*, 2023). However, significant challenges remain regarding their reliability, interpretability, alignment with clinical workflows, and potential for harm when deployed in high-stakes healthcare environments (Wiens *et al.*, 2019; Sendak *et al.*, 2020).

This year's session at the 2026 Pacific Symposium on Biocomputing (PSB), titled "AI and Machine Learning in Clinical Medicine: Bridging or Separating Model Intelligence and Human Expertise", examines the evolving relationship between computational capabilities and human clinical judgment in healthcare settings (Nateghi Haredasht, Kim, *et al.*, 2024). The papers in this session explore various facets of this complex relationship, representing cutting-edge research across four major themes: (1) LLMs in clinical applications, (2) AI for medical imaging analysis and interpretation, (3) approaches to clinical decision support and risk prediction, and (4) clinical data analysis and patient care. Here, we highlight the accepted submissions for this session that provide insights as healthcare institutions aim to integrate AI tools into clinical workflows while preserving the essential human dimensions of medical care.

2. Artificial Intelligence in Clinical Medicine

2.1. Large Language Models (LLMs) in Clinical Applications

The emergence of LLMs has expanded the possibilities for AI applications in clinical medicine. Several papers in this session explore how these powerful models can be adapted, evaluated, and deployed to address healthcare challenges.

One study investigating named entity recognition (NER) for substance use-related information found that fine-tuned encoder-based models consistently outperform LLMs in precision and span accuracy, highlighting a persistent gap between human expert knowledge and current AI capabilities for complex tasks requiring deep domain understanding (Dey and et al., 2026). Similarly, work on adverse drug event detection demonstrated that specialized models like RoBERTa Large achieved superior performance in identifying complex and contextually ambiguous events, with ensemble methods further improving relation extraction for pharmacovigilance applications (Prioleau and et al., 2026).

In clinical communication assessment, a couple of papers have developed frameworks leveraging LLMs to evaluate physician-patient interactions. One approach uses rubric-based Chain-of-Thought prompting with few-shot learning to assess the quality of risk communication in prostate cancer consultations, achieving expert-level agreement and establishing a scalable foundation for evaluating physician-patient communication in oncology (Lopez-Garcia and et al., 2026).

For clinical decision support and prediction, several papers explored novel applications of foundation models. Burkhart et al. developed methods to quantify the "informativeness" of clinical events in EHRs by measuring divergence between model predictions and observed outcomes, identifying "surprising" events that are predictive of adverse outcomes such as mortality (Burkhart and et al., 2026). However, when evaluated on emergency department revisit prediction tasks using comprehensive clinical data, LLM-based approaches underperformed traditional machine learning models that used only structured EHR data, suggesting that current LLM reasoning approaches may have limitations for certain clinical prediction tasks (Emma Chen and et al., 2026)

The session features several important evaluation frameworks and benchmarks. The ReXrank Challenge presents a benchmark for chest X-ray report generation models, revealing that while AI systems perform well on normal cases, they still struggle significantly with abnormal findings, demonstrating that automated radiology report generation remains largely unsolved for clinically significant pathologies (Zhang and et al., 2026). Another research team introduced MedAgentBench v2, advancing benchmark methodologies for evaluating and improving LLM-based agents in medical contexts through refined prompt engineering, enhanced EHR interaction tools, and novel memory mechanisms that allow agents to learn from past errors (E. Chen and et al., 2026).

For improving clinical documentation quality, Grolleau et al. developed MedFactEval, a framework for scalable, fact-grounded evaluation using an "LLM Jury" to assess AI-generated clinical summaries, achieving nearly perfect agreement with physician panels (Grolleau and et al., 2026). Another work on eConsult templates showed that while models like o3 demonstrated high comprehensiveness, they consistently struggled to prioritize the most clinically important questions, especially in narrative-heavy specialties (McCoy and et al., 2026). To address safety concerns, one team introduced a Retrieval-Augmented Error Checking pipeline that significantly improved error identification in AI-drafted patient-portal messages by incorporating local context retrieval (Chen and et al., 2026).

Across these diverse applications, we see both the potential and current limitations of LLMs in clinical medicine, with particular challenges remaining in achieving human-level domain expertise, prioritizing clinically relevant information, and ensuring factual accuracy in high-stakes healthcare applications.

2.2. Medical Imaging and Visual AI

Visual information processing represents another critical domain where AI is transforming clinical practice, from endoscopy to radiology to medical image interpretation. Research in this area focuses on developing AI systems that can enhance clinical workflows through improved visualization, interpretation, and analysis of medical imaging data.

In the field of endoscopic imaging, Hardy et al. have developed ColonCrafter, a diffusion-based depth estimation model designed to generate temporally consistent depth maps from monocular colonoscopy videos. Their approach leverages robust geometric priors learned from synthetic colonoscopy sequences while incorporating a novel style transfer technique to bridge the domain gap between synthetic training data and real clinical videos. This work enables important applications like improved 3D reconstruction of the colon and precise lesion localization, potentially enhancing the detection of subtle abnormalities during colonoscopy procedures (Hardy and et al., 2026).

Addressing the fragmentation in endoscopic AI development, Johri et al. introduce PanEndoAtlas, a clinician-guided dataset containing over 420,000 labeled endoscopic images compiled from 30 public datasets across 13 countries and 26 hospitals. This resource addresses the gap between current endoscopic AI capabilities and actual clinical needs. The team also presents PanEndoX, a benchmark suite of 10 clinically relevant tasks designed to evaluate the generalization capabilities of vision foundation models across the entire gastrointestinal spectrum, utilizing a hierarchical taxonomy for diagnostic reasoning that aligns with clinical practice (Johri and et al., 2026).

For musculoskeletal imaging, Sambara et al. have developed 3DReasonKnee, the first 3D grounded reasoning dataset specifically designed for medical image analysis. This resource comprises 494,000 quintuples derived from 7,970 3D knee MRI volumes, where each quintuple includes a 3D MRI volume, a diagnostic question, a 3D bounding box highlighting anatomical structures, clinician-generated reasoning steps, and structured severity assessments. This dataset serves as a testbed for advancing multimodal medical AI systems toward clinically aligned, localized decision-making in orthopedic imaging (Sambara and et al., 2026).

In chest radiology, Pal et al. present ReXVQA, a large-scale visual question answering benchmark featuring approximately 696,000 questions paired with 160,000 chest X-ray studies. Their work evaluates AI models on five core radiological reasoning skills: presence detection, location identification, negation understanding, differential diagnosis formulation, and geometric reasoning. Notably, their evaluation revealed that the best-performing AI model, MedGemma, achieved superior accuracy compared to human radiology residents on randomly sampled cases, suggesting progress in automated chest X-ray interpretation capabilities (Pal and et al., 2026).

These advances in medical imaging AI collectively demonstrate progress toward systems that not only perform well on technical metrics but also align with clinical workflows and reasoning processes.

2.3. Clinical Decision Support and Risk Prediction

AI systems show promise for supporting clinical decision-making and predicting patient risks across various healthcare domains. The papers in this section explore approaches to modeling complex medical phenomena, predicting adverse events, managing uncertainty, and enhancing clinical communication.

In the domain of infectious disease modeling, Liu et al. introduce EpiDHGNN, a Human Contact-Tracing Hypergraph Neural Network framework that leverages hypergraphs to capture intricate, higher-order relationships at both location and individual levels. By modeling the complex dynamics of human interactions, their approach outperforms baseline models in critical epidemic management tasks such as source detection and forecasting, offering potential improvements for public health response systems (Liu and et al., 2026).

For neurological applications, Feng et al. present SeizureFormer, a Transformer-based model for long-horizon seizure risk forecasting (1–14 days) using structured biomarkers derived from responsive neurostimulation (RNS) systems. Their approach integrates multi-scale CNN patch embedding, cross-variable temporal convolution, and squeeze-and-excitation attention mechanisms to capture both short-term fluctuations and long-term seizure cycles. By achieving state-of-the-art performance on this challenging prediction task, SeizureFormer demonstrates the potential to enable more proactive seizure management strategies for patients with drug-resistant epilepsy (Feng and et al., 2026).

Schmitz et al. explore the use of consumer-grade virtual reality (VR) headsets for automated neurological assessment, specifically focusing on Parkinson's disease classification from eye-tracking data. By combining general oculomotor metrics with task-evoked features and learned representations, their approach achieves high discriminative performance in both binary and multi-class classification scenarios, suggesting a cost-effective and accessible approach to neurodegenerative disease assessment (Schmitz and et al., 2026).

In psychiatric applications, Strobl introduces DEBIAS (Durable Effects with Backdoor-Invariant Aggregated Symptoms), an algorithm designed to learn causally predictable outcomes from psychiatric longitudinal data. By optimizing outcome definitions to maximize causal identifiability, this approach learns clinically interpretable weights for outcome aggregation while minimizing both observed and latent confounding. The algorithm outperforms state-of-the-art methods in recovering causal effects for composite outcomes in depression and schizophrenia, addressing a challenge in psychiatric outcomes research (Strobl and et al., 2026).

Addressing the critical issue of uncertainty in clinical AI systems, Ko et al. investigate abstention mechanisms for AI-based diagnostic classifiers, using pediatric autism video assessments as a case study. Their work demonstrates that carefully selecting upper and lower confidence thresholds can improve clinical performance metrics without requiring model retraining, offering a practical and interpretable approach to tailoring model behavior for diverse clinical contexts. This research highlights the importance of human-in-the-loop systems that can appropriately defer to human judgment in uncertain cases (Ko and et al., 2026).

Focusing on patient education, Luo et al. introduce ED-Explain, a system that transforms emergency department discharge instructions into personalized video presentations featuring a virtual healthcare provider. In evaluations by emergency physicians, these AI-generated video summaries received higher ratings for completeness, correctness, and patient accessibility compared to original text instructions (Luo and et al., 2026).

Finally, Ahsen et al. propose a computational framework that leverages generative AI to design interpretable structural causal models (SCMs) for clinical decision support. Their comparative analysis of transformer-based LLM against human experts on complex causal reasoning tasks revealed that both successfully designed clinically plausible causal diagrams with similar statistical performance (Ahsen and et al., 2026).

Collectively, these studies demonstrate how advanced AI techniques can enhance clinical decision-making across multiple dimensions: from modeling complex epidemiological patterns to forecasting individual patient risks, managing diagnostic uncertainty, improving patient communication, and providing interpretable decision support frameworks.

2.4. Clinical Data Analysis and Patient Care

Several papers in this session focus on methods for analyzing clinical data and improving patient care through AI, with particular emphasis on extracting meaningful information from unstructured data and evaluating AI systems in real-world clinical contexts.

Hwang et al. explore the application of LLMs for automating Multiple Sclerosis (MS) progression assessment from clinical notes. Their feasibility study examines LLMs' ability to derive Expanded Disability Status Scale (EDSS) and Functional Systems (FS) scores, metrics for monitoring MS progression. Their LLM-based solution demonstrated robust performance in classifying several FS subscores, particularly excelling in visual, sensory, and cerebellar domains, where it outperformed previous NLP systems. However, for overall EDSS classification, it did not exceed the performance of existing rules/CNN-based classifiers. This work highlights both the potential and current limitations of LLMs for extracting complex clinical assessments from unstructured text (Hwang and et al., 2026).

Pugh et al. introduce WATCH-SS (Warning Assessment and Alerting Tool for Cognitive Health from Spontaneous Speech), an innovative three-stage modular framework for detecting cognitive impairment from patients' spontaneous speech samples. This approach addresses the need for explainable AI in healthcare by leveraging detectors for five linguistic and acoustic indicators of cognitive impairment, aggregating their outputs into clinically interpretable features, and employing a predictive model for classification. The framework achieved a performance of AUC = 0.80 (Pugh and et al., 2026).

Addressing the crucial issue of real-world AI performance, Banerjee et al. present the ReXecution framework for clinician-centered assessments of medical AI assistants. Their evaluation of AI systems for chest X-ray interpretation in realistic clinical settings revealed a significant "intention-execution disconnect"—despite demonstrating considerable medical knowledge, ChatGPT-o3 and MedGemma frequently struggled to accurately interpret images and execute tasks, producing correct outputs in only 5-10% of cases. This stark finding highlights the gap between AI system capabilities in controlled evaluation settings versus real-world clinical deployment, underscoring the need for more robust evaluation frameworks that reflect actual clinical usage scenarios (Banerjee and et al., 2026).

Wu et al. tackle the challenge of evaluating AI system performance in clinical consultation contexts by developing automated methods to assess concordance between AI-generated and human specialist responses in physician-to-physician electronic consultations (eConsults). Their study compared two approaches, "LLM-as-judge" and "decompose-then-verify", finding that the "LLM-as-judge" method achieved human-level concordance assessment with an F1-score of 0.89 and Cohen's κ of 0.75, comparable to inter-physician agreement (Wu and et al., 2026).

Zolensky et al. introduce methods for speaker role identification in clinical conversations, addressing a challenge in the automated analysis of doctor-patient interactions. Their work demonstrates techniques to identify different speakers and their roles in clinical dialogue, providing foundational capabilities for subsequent analyses of communication patterns, information exchange, and relationship dynamics in healthcare encounters (Zolensky and et al., 2026).

Finally, Chen et. al. evaluate multiple NLP techniques for classifying Post-Traumatic Stress Disorder (PTSD) from clinical interview transcripts. The study compares embedding-based methods (like SentenceBERT), fine-tuned transformers (like Mental-RoBERTa), and various LLM prompting strategies (zero-shot, few-shot). Key findings show that a SentenceBERT embedding model with logistic regression achieved the best overall performance, outperforming more complex models, and that few-shot LLM prompting with DSM-5 criteria also yielded competitive results (F. Chen and et al., 2026).

Collectively, these papers illustrate diverse approaches to leveraging AI for clinical data analysis and patient care improvement, while emphasizing the importance of explainability, trustworthiness, and realistic performance evaluation in healthcare AI systems. They highlight both significant progress and persistent challenges in developing AI tools that can effectively support clinical practice.

3. Conclusion

The papers presented in this session demonstrate significant progress in applying AI to clinical medicine while highlighting the fundamental challenge of creating systems that complement rather than replace human expertise. Three key themes emerge: the growing importance of multimodal reasoning capabilities that integrate diverse data sources; the persistent need for interpretability and trustworthiness in increasingly complex systems; and the critical importance of evaluating AI performance in realistic clinical contexts.

Looking forward, promising directions include the integration of causal reasoning, enhanced multimodal understanding, and robust uncertainty quantification methods. The central question posed by our session title—whether AI is bridging or separating model intelligence and human expertise—remains nuanced, but the most promising path appears to be designing AI systems that deliberately complement human capabilities. Such systems would address clinicians' cognitive limitations while leveraging their irreplaceable judgment, empathy, and contextual understanding. The research presented here represents meaningful progress toward achieving productive human-AI collaboration in healthcare that enhances rather than diminishes the role of clinical expertise.

References

Ahsen and et al. (2026) "Leveraging Generative AI for Interpretable Clinical Decision Making Through Causal Graphs," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Ashtari, P. *et al.* (2022) "New multiple sclerosis lesion segmentation and detection using pre-activation U-Net," *Frontiers in Neuroscience*, 16. Available at: https://doi.org/10.3389/fnins.2022.975862.

Ashtari, P. *et al.* (2023) "Factorizer: A scalable interpretable approach to context modeling for medical image segmentation," *Medical Image Analysis*, 84, p. 102706. Available at: https://doi.org/10.1016/j.media.2022.102706.

Ashtari, P., Maes, F. and Van Huffel, S. (2021) "Low-Rank Convolutional Networks for Brain Tumor Segmentation," in A. Crimi and S. Bakas (eds.) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer International Publishing, pp. 470–480. Available at: https://doi.org/10.1007/978-3-030-72084-1_42.

Banerjee and et al. (2026) "The Intention-Execution Disconnect in Medical AI: The ReXecution Framework," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Bedi, S. *et al.* (2025) "MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks." arXiv. Available at: https://doi.org/10.48550/arXiv.2505.23802.

Bommasani, R. *et al.* (2022) "On the Opportunities and Risks of Foundation Models." arXiv. Available at: https://doi.org/10.48550/arXiv.2108.07258.

Burkhart and et al. (2026) "Quantifying surprise in clinical care with foundation models," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Chang, C.T. *et al.* (2025) "Red teaming ChatGPT in medicine to yield real-world insights on model behavior," *npj Digital Medicine*, 8(1), p. 149. Available at: https://doi.org/10.1038/s41746-025-01542-0.

Chen, E. and et al. (2026) "MedAgentBench v2: Improving Medical LLM Agent Design," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Chen and et al. (2026) "Retrieval-Augmented Guardrails for AI-Drafted Patient-Portal Messages," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Chen, F. and et al. (2026) "Detecting PTSD in Clinical Interviews: Comparative Analysis of NLP Methods and LLMs," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Davenport, T. and Kalakota, R. (2019) “The potential for artificial intelligence in healthcare,” *Future Healthcare Journal*, 6(2), pp. 94–98. Available at: https://doi.org/10.7861/futurehosp.6-2-94.

Dey and et al. (2026) “Inference Gap in Domain Expertise and Machine Intelligence in Named Entity Recognition,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Emma Chen and et al. (2026) “Turning Large Language Models into Emergency Department Revisit Predictors,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Feng and et al. (2026) “SeizureFormer: A Multi-Scale Transformer for Seizure Risk Forecasting,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Grolleau and et al. (2026) “MedFactEval and MedAgentBrief: Framework for Generating and Evaluating Factual Clinical Summaries,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Hardy and et al. (2026) “ColonCrafter: A Depth Estimation Model for Colonoscopy Videos Using Diffusion Priors,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Hwang and et al. (2026) “Leveraging Large Language Models to Derive Multiple Sclerosis Progression Assessments from Clinical Notes,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Johri and et al. (2026) “A Clinician-Guided Framework for Endoscopic AI: Developing PanEndoAtlas,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Ko and et al. (2026) “Abstention and Threshold Identification for Uncertainty Management in Clinical Decision Tools,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Liu and et al. (2026) “Higher-order Interaction Matters: Modeling Epidemics via Dynamic Hypergraph Neural Networks,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Lopez, I. *et al.* (2025) “Clinical entity augmented retrieval for clinical information extraction,” *npj Digital Medicine*, 8(1), pp. 1–11. Available at: https://doi.org/10.1038/s41746-024-01377-1.

Lopez-Garcia and et al. (2026) “Scoring Physician Risk Communication in Prostate Cancer Using Large Language Models,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Luo and et al. (2026) “ED-Explain: Personalized Video Instructions for Patients Discharged from the Emergency Department,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

McCoy and et al. (2026) “Asking the Right Questions: Benchmarking LLMs in Clinical Consultation Templates,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Nateghi Haredasht, F. *et al.* (2023) “Predicting outcomes of acute kidney injury in critically ill patients using machine learning,” *Scientific Reports*, 13, p. 9864. Available at: https://doi.org/10.1038/s41598-023-36782-1.

Nateghi Haredasht, F., Fouladvand, S., *et al.* (2024) "Predictability of buprenorphine-naloxone treatment retention: A multi-site analysis combining electronic health records and machine learning," *Addiction*, 119(10), pp. 1792–1802. Available at: https://doi.org/10.1111/add.16587.

Nateghi Haredasht, F., Kim, D., *et al.* (2024) "Session Introduction: AI and Machine Learning in Clinical Medicine: Generative and Interactive Systems at the Human-Machine Interface," in *Biocomputing 2025*. WORLD SCIENTIFIC, pp. 33–39. Available at: https://doi.org/10.1142/9789819807024_0003.

Nateghi Haredasht, F. and Vens, C. (2022) "Predicting Survival Outcomes in the Presence of Unlabeled Data," *Machine Learning*, 111(11), pp. 4139–4157. Available at: https://doi.org/10.1007/s10994-022-06257-x.

Noei, S. *et al.* (2016) "Classification of EEG signals using the Spatio-temporal feature selection via the elastic net," in *2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME). 2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME)*, pp. 232–236. Available at: https://doi.org/10.1109/ICBME.2016.7890962.

Nori, H. *et al.* (2023) "Capabilities of GPT-4 on Medical Challenge Problems." arXiv. Available at: https://doi.org/10.48550/arXiv.2303.13375.

Pal and et al. (2026) "ReXVQA: A Large-scale Visual Question Answering Benchmark for Chest X-ray Understanding," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Patel, B.N. *et al.* (2019) "Human–machine partnership with artificial intelligence for chest radiograph diagnosis," *npj Digital Medicine*, 2(1), p. 111. Available at: https://doi.org/10.1038/s41746-019-0189-7.

Prioleau and et al. (2026) "Leveraging Large Language Models for Adverse Drug Event Detection," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Pugh and et al. (2026) "WATCH-SS: A Trustworthy and Explainable Modular Framework for Detecting Cognitive Impairment," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Rajkomar, A., Dean, J. and Kohane, I. (2019) "Machine Learning in Medicine," *The New England Journal of Medicine*, 380(14), pp. 1347–1358. Available at: https://doi.org/10.1056/NEJMra1814259.

Sambara and et al. (2026) "3DReasonKnee: Advancing Grounded Reasoning in Medical Vision Language Models," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Schmitz and et al. (2026) "Towards Automated Analysis of Gaze Behavior from Consumer VR Devices for Neurological Diagnosis," in *Pacific Symposium on Biocomputing (PSB 2026)*.

Sendak, M.P. *et al.* (2020) “A Path for Translation of Machine Learning Products into Healthcare Delivery,” *European Medical Journal* [Preprint]. Available at: https://doi.org/10.33590/emjinnov/19-00172.

Strobl and et al. (2026) “Learning Causally Predictable Outcomes from Psychiatric Longitudinal Data,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Thirunavukarasu, A.J. *et al.* (2023) “Large language models in medicine,” *Nature Medicine*, 29(8), pp. 1930–1940. Available at: https://doi.org/10.1038/s41591-023-02448-8.

Wiens, J. *et al.* (2019) “Do no harm: a roadmap for responsible machine learning for health care,” *Nature Medicine*, 25(9), pp. 1337–1340. Available at: https://doi.org/10.1038/s41591-019-0548-6.

Wu and et al. (2026) “Automated Evaluation of Large Language Model Response Concordance with Human Specialist Responses on eConsult Cases,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Zhang and et al. (2026) “Automated Chest X-ray Report Generation Remains Unsolved,” in *Pacific Symposium on Biocomputing (PSB 2026)*.

Zolensky and et al. (2026) “Speaker Role Identification in Clinical Conversations,” in *Pacific Symposium on Biocomputing (PSB 2026)*.