AI and Machine Learning in Clinical Medicine: From Research to Real-World Deployment

Fateme Nateghi Haredasht[1,*], David Wu[2], David JH Wu[3], Austin Schoeffler[4], Kameron C. Black[5], Dokyoon Kim[6], Joseph Romano[6], Geoffrey Tison[7], Brett Beaulieu-Jones[8]

[1]*Stanford Division of Computational Medicine, Stanford University, Stanford, CA, USA*
[2]*Department of Dermatology, Mass General Brigham, Harvard Medical School, Boston, MA, USA*
[3]*Department of Radiation Oncology, Stanford Cancer Center, Palo Alto, CA, USA*
[4]*Department of Emergency Medicine, Stanford University School of Medicine, Stanford, CA*
[5]*Division of Hospital Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA*
[6]*Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[7]*Division of Cardiology and Center for Biosignal Research, University of California, San Francisco, San Francisco, CA, USA*
[8]*Section of Computational Biomedicine and Biomedical Data Science, Department of Medicine, University of Chicago, Chicago, IL, USA*
[*]*Corresponding author. Email: fnateghi@stanford.edu*

Artificial intelligence (AI) in clinical medicine is moving from isolated prediction tasks toward systems that retrieve evidence, combine modalities, interact with users, and operate in care workflows. Aggregate benchmark success can coexist with subgroup failures, site or representation shifts, multi-turn errors, and downstream decision failures. This session examines three themes: clinically meaningful failure modes; representation, harmonization, and multimodal transfer; and deployment pipelines that preserve evidence for audit and human review. Across them, clinical validity must be established for every component, mapping, and interface in its intended use.
Keywords: Artificial intelligence; clinical medicine; clinical validation; multimodal learning; large language models; real-world deployment.

1. Introduction

Machine learning supports medical imaging, risk prediction, clinical natural language processing, and decision support. For computational biology, translation means moving beyond research-dataset predictions toward useful clinical action. Performance depends on data acquisition and representation, concept mapping, population-specific errors, user responses, and downstream system consequences (Rajkomar, Dean and Kohane, 2019; Wu *et al.*, 2025; Goh *et al.*, 2026).

Foundation models and large language models (LLMs) widen this gap because behavior depends on prompts, evidence, conversational context, and perceived user expertise; multimodal systems add dependencies among encoders, fusion, and downstream tasks. The deployable object is therefore a workflow-embedded configuration, not an isolated model. This extends prior sessions' emphasis on explainability and generalizability (Nateghi Haredasht et al., 2024) and on aligning model intelligence with human expertise (Nateghi Haredasht et al., 2025).

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

The 2027 PSB session, "AI and Machine Learning in Clinical Medicine: From Research to Real-World Deployment," brings together ten papers spanning diagnostic dialogue, precision oncology, neuroimaging, pathology, multimodal learning, information extraction, and reinforcement learning. Together, they ask what evidence makes a technical gain trustworthy in its intended clinical setting.

2. Artificial Intelligence in Clinical Medicine

2.1. Auditing clinically meaningful failure modes

Kim et al. evaluate three frontier LLMs in 897 paired four-turn diagnostic conversations with lay-patient and specialist interlocutors. Models revised diagnoses for specialists 16–35 times more often than for patients; 74%–83% of revisions from an initially correct diagnosis ended incorrectly. Content-matched controls isolated specialist register from clinical content introduced during questioning (Kim et al., 2027).

In a 500-case synthetic non-small cell lung cancer benchmark, Ashrafi et al. found that eight LLM configurations misclassified 83.3%–100% of Partially Supported recommendations. MTB-AuditAgent reduced over-refusal to 6.7% and achieved 91.2% accuracy (95% CI, 88.6%–93.6%). Disagreements between two oncologists centered on information sufficiency and treatment optimization, showing that clinically meaningful safety distinctions require explicit definitions (Ashrafi et al., 2027).

Across 100 liver-biopsy whole-slide images from two centers, Yang et al. found close agreement between AI-derived lipid-droplet area and expert measurement. Yet a Center A mapping from area to historical NASH CRN steatosis grade lost substantial agreement at Center B, as did expert-derived area, locating the failure in the measurement-to-grade mapping rather than segmentation (Yang et al., 2027).

Together, these studies show how headline metrics can conceal clinically important error. Expert review supports calibration in tumor board auditing and pathology, yet expert-like interaction can trigger deference. Oversight must therefore follow new evidence and explicit decision rules, not authority alone.

2.2. Representation, harmonization, and multimodal transfer

Chattopadhyay et al. compare CycleGAN and voxelwise ComBat harmonization across four dementia imaging cohorts and four prediction tasks. Stronger image similarity did not consistently improve clinical prediction; results depended on endpoint, the diagnostic mix used to fit the harmonizer, and applying the same transformation during training and inference (Chattopadhyay et al., 2027).

Nguyen and Yang hold three vision-language decoders fixed while varying four retrieval encoders for mammography report generation. Only the mammography-native encoder improved

held-out retrieval after contrastive fine-tuning (precision at one, 0.513 to 0.585). Gains concentrated in routine screening, while biopsy-recommendation accuracy remained poor; aggregate improvement did not establish transfer to the most consequential stratum (Nguyen and Yang, 2027).

Golovanevsky et al. test all 26 combinations of five MIMIC modalities and show that adding modalities can degrade contrastive alignment and transfer. A modality-gated LSTM improved mortality AUROC from 73.19% to 76.93% and AUPRC from 51.27% to 62.26% over the five-modality supervised baseline. Evaluation remains internal to complete-case MIMIC data (Golovanevsky et al., 2027).

Yu et al. develop a TabPFN concept bottleneck grounded in nine Alzheimer disease biomarkers. In 505 ADNI participants, mean AUC was 0.764, near direct TabPFN's 0.766. Simulated correction of CDR Sum of Boxes raised AUC from 0.765 to 0.943, but replacing every predicted concept reduced performance, exposing distribution shift between predicted and measured concepts (Yu et al., 2027).

Together, these papers show that better alignment, embeddings, additional modalities, or intermediate concepts do not guarantee better clinical outputs. Validation must follow the path from representation to decision and report results for groups in which errors have greatest consequence.

2.3. From model output to accountable workflow

Gupta et al. describe an extraction system deployed across four HIV and tuberculosis programs in South Africa and Zimbabwe, serving over 50,000 users. A 398-conversation benchmark selected models; 134 scenario tests gated model-prompt-schema-validator configurations. Similar aggregate scores concealed important HIV-status errors, supporting deterministic checks wherever expected records could be specified (Gupta et al., 2027).

Ma et al. combine page-preserving parsing, temporal selection, constrained LLM extraction, and source-page links for scanned longitudinal records. Masked physician adjudication found 626 correct fields among 628 comparisons across 157 heart-transplant packages, versus 518 for an external abstraction service. Both AI errors arose in text recovery; independent adjudication and prospective evaluation remain needed (Ma et al., 2027).

Li et al. compare set-valued policies retaining near-optimal actions with dead-end discovery excluding actions above an absolute risk threshold. In simulation and a 19,626-patient MIMIC-III sepsis cohort, conflicts were uncommon but concentrated in severe states. Their partial ordering exposes conflict and preserves clinician judgment without claiming a validated treatment policy (Li, Wu and Tang, 2027).

These papers span a research-to-deployment gradient: an operating service, internal record validation, and retrospective testing of decision-support logic. All make provenance, release criteria, and unresolved conflict visible rather than treating them as implementation details.

3. Conclusion

This year's papers ask whether the full system remains valid as evaluation approaches practice. Three lessons recur: target asymmetric failures and consequential subgroups, not only aggregates; test validity across sites, representations, modalities, and downstream mappings; and require traceable evidence, release gates, and bounded roles for human judgment.

"Real-world deployment" is not one threshold. Evidence spans retrospective experiments, external-site tests, internal workflow validation, and an operating multi-country service. Progress requires naming these stages and matching claims to evidence. The strongest path is not autonomy by default, but systems whose components, limitations, and human responsibilities can be examined under conditions of care.

References

Ashrafi, N., Luo, J., Frumm, S.M. and Daneshjou, R. (2027) "OpenMTB-Audit: Exposing Over-Refusal and Clinical Expert Perspectives in LLM-Based Molecular Tumor Board Safety Evaluation," in *Pacific Symposium on Biocomputing (PSB 2027)*.

Chattopadhyay, T., Mehendale, G., Ankarath, R.H., Thomopoulos, S.I., Joshi, H., Venkatasubramanian, G., John, J.P., Ambite, J.L., Ver Steeg, G. and Thompson, P.M. (2027) "Evaluating MRI Harmonization via Multiple Clinical Prediction Tasks," in *Pacific Symposium on Biocomputing (PSB 2027)*.

Goh, E. *et al.* (2026) "Toward a test of medical AI superintelligence," *Nature Medicine*, 32(9), pp. 3128–3130. Available at: https://doi.org/10.1038/s41591-026-04539-8.

Golovanevsky, M., Mahableshwarkar, P., Singh, R. and Eickhoff, C. (2027) "PiCME: Pipeline for Contrastive Modality Evaluation and Encoding in the MIMIC Dataset," in *Pacific Symposium on Biocomputing (PSB 2027)*.

Gupta, K., Mbatuegwu, D., Mwongela, S., Richard, Y. and Mendonca, R. (2027) "A Deployed Framework for Structured Clinical Extraction from Multi-Turn Health Conversations: From Model Selection to Scenario-Based Release Gating," in *Pacific Symposium on Biocomputing (PSB 2027)*.

Kim, T.M., Luo, L., Kim, S.E., Willauer, A.N. and Rajpurkar, P. (2027) "The White Coat Effect: Specialist Personas Drive Harmful Diagnostic Drift More Than Patient Personas in Clinical LLMs," in *Pacific Symposium on Biocomputing (PSB 2027)*.

Li, V., Wu, S. and Tang, S. (2027) "Reconciling Set-Valued Policy & Dead-End Discovery in Healthcare Reinforcement Learning: An Empirical Analysis," in *Pacific Symposium on Biocomputing (PSB 2027)*.

Ma, M., Tao, A., Focht, C., Huang, J., Teuteberg, J., Kobulnik, J. and Li, L. (2027) "Temporally Grounded and Evidence-Linked Clinical Abstraction from Scanned Longitudinal Records in Heart Transplant Recipients," in *Pacific Symposium on Biocomputing (PSB 2027)*.

Nateghi Haredasht, F., Kim, D., Romano, J.D., Tison, G., Daneshjou, R. and Chen, J.H. (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., Romano, J.D., Beaulieu-Jones, B.K., Kim, D., Ioannidis, A., Tison, G.H., Daneshjou, R. and Chen, J.H. (2025) "Session Introduction: AI and Machine Learning in Clinical Medicine Bridging or Separating Model Intelligence and Human Expertise," in *Biocomputing 2026*. World Scientific, pp. 1–11. Available at: https://doi.org/10.1142/9789819824755_0001.

Nguyen, D. and Yang, W. (2027) "Separation Without Transfer: The Retrieval Encoder's Substrate, Not the VLM Decoder, Decides Whether Contrastive Fine-Tuning Transfers in Mammography Report Generation," in *Pacific Symposium on Biocomputing (PSB 2027)*.

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.

Wu, D. *et al.* (2025) "First, do NOHARM: towards clinically safe large language models," *arXiv* [Preprint]. Available at: https://doi.org/10.48550/arXiv.2512.01241.

Yang, S.R., Wu, T.H., Tsai, H.W., Chiu, H.C., Hsu, C.W., Hsu, W.L., Chung, P.C. and Alterovitz, G. (2027) "Measurement Is Not the Grade: A Two-Center Whole-Slide Study of AI-Based Lipid-Droplet Area Quantification and NASH CRN Steatosis Grading," in *Pacific Symposium on Biocomputing (PSB 2027)*.

Yu, T., Yang, S., Wen, Z., Tong, B., Han, T., Davatzikos, C., Shen, L. and the Alzheimer's Disease Neuroimaging Initiative (2027) "TabPFN-CBM: A Concept-Interpretable Tabular Foundation Model Framework for Multimodal Alzheimer's Diagnosis," in *Pacific Symposium on Biocomputing (PSB 2027)*.