independently. This study leverages computer vision algorithms to create a scalable, low-cost, and accessible tool that automatically quantifies SMC behavior in infants born preterm, allowing for earlier screening and intervention for CP. METHODS/STUDY POPULATION: In single-camera videos of 600 15-week-old infants, all of whom have standardized clinical SMC scores available, human pose estimation algorithms will be used to derive 3D joint trajectories and kinematic metrics. These features will be used to train and cross-validate a machine learning model aimed at automating clinical scoring of SMC from single-camera video recordings of infants. Then, this model will be used to automatically score the SMC in longitudinal single-camera videos of 47 preterm infants (collected every two weeks from birth until 16 weeks of age). The SMC score trajectories between infants who eventually received a CP diagnosis and those who did not will be compared to establish an inflection point at which SMC development begins to diverge, establishing an optimal timepoint for intervention. RESULTS/ANTICIPATED RESULTS: Preliminary results show that 3D angles of 22 joints across the infant's body can be estimated in each frame of a 1 to 3 minute single-camera video. We anticipate that the error between the 3D angle estimated from a custom fitted kinematic model and the 2D angle calculated using human pose estimation measures will be less than 5 degrees for all joints measured. With a dataset of over 1,500 videos from over 600 infants, we expect that the trained machine learning model will achieve a sensitivity and specificity of over 90% in predicting the clinical SMC score. Finally, we expect that the automated, longitudinal SMC score trajectories will begin to significantly deviate in infants with a later CP diagnosis, from those without, at around term age (40 weeks gestational age), based on recent clinical findings. DISCUSSION/SIGNIFICANCE OF IMPACT: We provide the first automated, quantitative analysis of SMC from video recordings that can be used within real-world settings such as the NICU, at home, or at daycare. This work will result in a cost-effective, precise, smartphone-accessible tool that empowers earlier, more equitable diagnosis and treatment planning.

57

Translational evaluation of multimodal artificial intelligence for dermatology triage

Nuran Golbasi[1], Olivier Elemento[2], Veronica Rotemberg[3] and Shari R. Lipner[4]

[1]Weill Cornell Medical College, New York; [2]Englander Institute for Precision Medicine, Institute for Computational Biomedicine, Weill Cornell Medicine, New York; [3]Dermatology Service, Memorial Sloan Kettering Cancer Center, New York and [4]Department of Dermatology, Weill Cornell Medicine, New York

OBJECTIVES/GOALS: To evaluate the translational reliability, reproducibility, diagnostic performance, and subgroup equity of multimodal artificial intelligence (AI) models for dermatology triage across multiple model platforms. METHODS/STUDY POPULATION: Limited access to dermatology expertise delays diagnosis and care, motivating development of multimodal AI systems that integrate clinical images with patient data for triage. We assembled 200 biopsy-confirmed PAD-UFES-20 lesions (melanoma, keratinocyte carcinoma, benign) with paired images and metadata, prioritizing demographic balance. Six multimodal AI models (GPT-5, GPT-5-mini; Gemini 2.5 Pro, Gemini 2.5 Flash; Claude Sonnet-4, Claude Opus-4) analyzed these lesions with identical prompts predicting diagnostic probabilities, triage (urgent vs routine), and rationale. Outcomes included sensitivity, specificity, AUROC, F1, and subgroup equity. Model rationales were reviewed for interpretability, and subset re-prompting tested reproducibility for translational robustness. RESULTS/ANTICIPATED RESULTS: Across six models, sensitivity range was 0.89–1.00, specificity 0.21–0.65, AUROC 0.77–0.87, and F1 scores 0.72–0.81. GPT-5 achieved the most balanced performance (0.92 sensitivity, 0.65 specificity, AUROC 0.87, F1 0.81), while Gemini 2.5 Pro and Flash reached perfect sensitivity but low specificity (0.21–0.25). Claude Sonnet-4 showed near-perfect sensitivity (0.99) but over-called benign cases (0.24 specificity), while Opus-4 had the lowest sensitivity (0.89). Urgent triage aligned with dermatologist biopsy patterns (87–97%), and sensitivity was consistent across sex and skin type ($p \geq 0.29$). Subset re-prompting produced similar results, supporting reproducibility. Model rationales reflected dermatologic reasoning, supporting interpretability, and translational readiness. DISCUSSION/SIGNIFICANCE OF IMPACT: Multimodal AI models showed balanced diagnostic performance for dermatology triage, with platform-specific trade-offs between sensitivity and specificity. Subgroup equity, interpretable rationales, and subset reproducibility define key elements for reliable translation into dermatology workflows and prospective validation.

60

Post-deployment artificial intelligence model monitoring, evaluation, and intervention in health systems: A scoping review for guidelines for AI model report and the literature

Chenyu Gai[1], Shauna M. Overgaard[2], Joshua W. Ohde[2], Momin M. Malik[2], Hanyin Wang[2], Matthew J. Spiten[2], Deepak K. Sharma[2], Chung Il Wi[2] and Young Juhn[2]

[1]Mayo Clinic and [2]Mayo Clinic, Rochester, Minnesota, USA

OBJECTIVES/GOALS: This scoping review aims to synthesize current literature on post-deployment monitoring of AI-enabled digital health solutions within clinical practice. Findings identify existing approaches and gaps that inform guidance for post-deployment monitoring in clinical practice. METHODS/STUDY POPULATION: We conducted a scoping review in accordance with PRISMA-ScR guidelines to characterize the current landscape of post-deployment monitoring in healthcare systems. A PubMed search targeted peer-reviewed articles in English published between 2015 and mid-2025, including text or MeSH terms on 1) health system/hospital; 2) artificial intelligence; 3) post-deployment; and 4) evaluation/monitoring. We performed a thematic analysis to identify common challenges, gaps, and opportunities in AI oversight. Additionally, we reviewed guidelines addressing post-deployment AI monitoring. All analyses were conducted using Rayyan.ai and Microsoft Word. RESULTS/ANTICIPATED RESULTS: Among the six studies included after the full-text review, five provide recommendations to ensure transparency, safety, and model performance.

https://doi.org/10.1017/cts.2026.10286 Published online by Cambridge University Press