A Multi-Agent Artificial Intelligence Model for ACL Surgery Evidence Synthesis

Purpose/Hypothesis: The aim was to develop a multi-agent AI system that automates retrieval, appraisal, and synthesis of anterior cruciate ligament (ACL) surgery evidence to generate clinically relevant, source-linked summaries for orthopedic surgeons. We hypothesized that the multi-agent pipeline would (1) retrieve and prioritize ACL literature with high factual precision and (2) generate coherent summaries that remain concordant with major ACL guideline recommendations.

Methods: We built a multi-agent system powered by language models (Llama-3.3-70B and Gemma2-9B) with agents for PubMed search, full-text retrieval, data extraction, evidence synthesis, and quality validation. The model was evaluated on 15 clinical questions, using recommendations derived from five ACL guidelines (AAOS, ESSKA, ISAKOS, IOC) as standards. Performance was assessed with automated metrics quantifying: (1) clinical entity extraction by F1 score, (2) quantitative accuracy, (3) contradiction detection using natural language inference, and (4) semantic concordance using Sentence-BERT cosine similarity between AI-generated summaries and guideline recommendations.

Results: Mean factual accuracy was 82% (range: 58-100%), with 95% entity extraction F1 score (range: 69-100%; 67% of questions achieved perfect extraction). Six questions (40%) achieved near-perfect overall accuracy (>98%). The system produced zero contradictions (0/15) with guideline recommendations. Semantic concordance was 78% (range: 59-89%). Quantitative accuracy was 60% (range: 0-100%), exhibiting a bimodal distribution with either excellent (>90%, 47% of questions) or poor (<25%, 20% of questions) performance, identifying a specific improvement target.

Conclusion: This ACL-focused multi-agent AI system achieved 82% factual accuracy with zero contradictions when synthesizing current ACL evidence. The model maintained guideline concordance while incorporating literature published years after guideline development, demonstrating potential for automated, real-time clinical evidence synthesis to supplement guideline recommendations between formal updates.