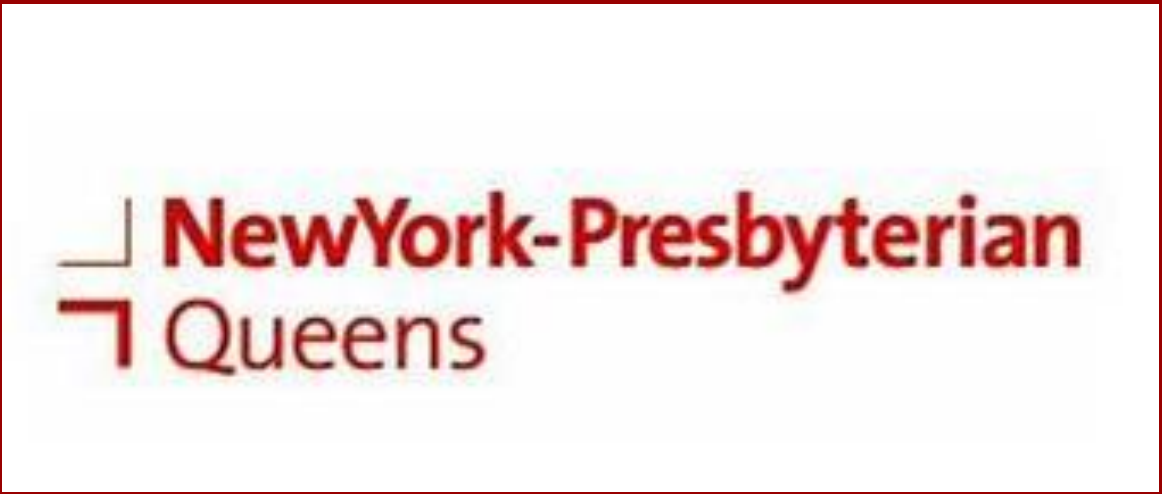




ARTIFICIAL INTELLIGENCE FOR EMERGENCY SURGERY DECISIONS: ACCURACY ON REAL CONSULT NOTES FROM THE ED

Phillip Hwang, MD; Catherine Qiu, MD; Daniella Abramov, DO; Durganili Balasubramaniyan, DO; Julie Hong, MD; Amber Chen-Goodspeed, MD; Steven Chao, MD, FACS



BACKGROUND

- Artificial intelligence may assist surgical decision-making.
- Domain-specific and general AI models may perform differently in clinical use.
- Emergency general surgery consults require accurate decisions on operative need.
- Real, de-identified consult notes allow evaluation in a practical setting.
- However, concerns remain regarding:
 - . Decision accuracy
 - . Diagnostic performance
 - . Reliability across EGS conditions

Study Aim

To compare OpenEvidence and Copilot using de-identified EGS consult notes, assessing accuracy of the “surgery indicated this admission” decision and diagnostic performance across common EGS conditions.

METHODS

- Design:** Single-center, retrospective paired diagnostic-accuracy study
- Sample:** Adult EGS consults (N=111)
- o 101 operative
 - o 10 non-operative
- Input:** Initial consult data only
- o HPI, vitals, exam
 - o Key labs, radiology
 - o De-identified
 - o Assessment/plan excluded
- Model Tasks:**
- o Primary diagnosis + 2-item differential
 - o “Surgery indicated this admission?” (Yes/No)
 - o If Yes, recommended operation
 - o ERCP/IR/bedside procedures not counted
- Comparison Standard:**
- o Attending diagnosis
 - o Actual index-admission management in EHR

RESULTS

- Surgery indicated (Yes/No):**
- OpenEvidence: TP 73, FP 3, TN 7, FN 28; Sensitivity 72%, Specificity 70%, Accuracy 72%.
 - Copilot: TP 96, FP 5, TN 5, FN 5; Sensitivity 95%, Specificity 50%, Accuracy 91%.
- Diagnostic performance (operative subgroups):**
- Laparoscopic cholecystectomy (n=46): OpenEvidence 80%, Copilot 87% correct diagnosis.
 - Laparoscopic appendectomy (n=37): OpenEvidence 91%, Copilot 95% correct diagnosis.

DISCUSSION/CONCLUSION

- Both systems produced usable outputs for the “surgery this admission” decision.
- However:**
- . Copilot showed higher sensitivity and overall accuracy.
 - . OpenEvidence showed a more balanced sensitivity and specificity.
 - . Accuracy was higher for common EGS diagnoses and lower for nuanced cases.
- Limitation:**
- . Surgery consult notes often already suggested diagnosis and operative need.
 - . Future studies should use ED or other specialty notes to test generalizability.
 - AI may support consult residents in diagnosis and triage.
 - . Use should remain supervised and adjunctive to clinical judgment.

For questions please contact: Phillip Hwnag MD PGY5
philliphwang823@gmail.com