BEYOND HUMAN CODING: COMPARING HUMAN EXPERTS VS ARTIFICIAL INTELLIGENCE IN INTERPRETING SURGEON-PROVIDED FEEDBACK DURING RESIDENT-PERFORMED LAPAROSCOPIC CHOLECYSTECTOMY

Rui-Wamba J, MD; Vildósola C; Becker B, MD; Aguilera M, MD; Castiglioni E, MD; Gaete M, MD, MSc; Jarry C, MD, MSc; Varas J, MD, MSc; Pontificia Universidad Católica de Chile

INTRODUCTION

• Structured feedback after resident-performed laparoscopic cholecystectomy can reveal recurrent technical errors and teachable micro-skills.

• Human coding remains the educational gold standard, but it is labor-intensive and difficult to scale.

• Large language models may enable rapid, taxonomy-based analysis of narrative faculty feedback.

OBJECTIVE

• Compare qualitative concordance and complementarity between human experts and ChatGPT when interpreting surgeon-provided feedback during resident-performed cholecystectomy.

20 PGY-1 residents	**46** Procedures
298 Coded errors	**34%** Without Critical View of Safety (CVS)

WHY THIS MATTERS

Scalable analysis could convert narrative comments into curriculum-level signals without replacing expert review.

METHODS

•Methodological comparative study of 46 laparoscopic cholecystectomies performed by 20 residents.

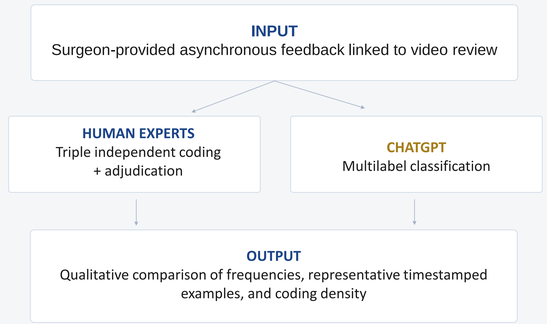
• Source material: asynchronous expert feedback linked to video review.

• Human pathway: triple independent coding with adjudication.

• AI pathway: ChatGPT multilabel classification using a predefined surgical taxonomy.

• Compared macro-categories, timestamped representative examples, and coding density.

COMPARATIVE CODING PIPELINE



DOMAINS COMPARED

Exposure, sequence/planes, hook/energy use, CVS, clipping, hemostasis, ergonomics, and related technical themes.

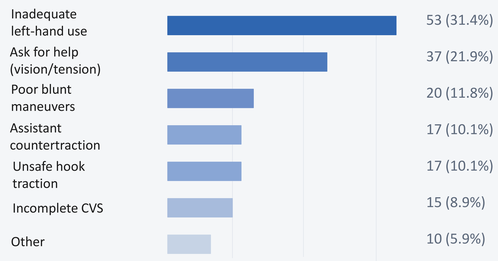
RESULTS

• Substantial qualitative convergence in the most frequent domains.

• Both approaches prioritized exposure, sequence/planes of dissection, and safe hook technique.

• Shared practical patterns: hemostasis before extraction, confirmation of two clip tips, and consistent completion of CVS.

MOST FREQUENT AI-CODED FEEDBACK DOMAINS



COMPLEMENTARY ANALYTIC STRENGTHS

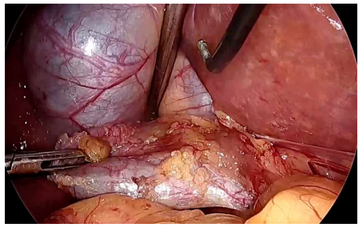
HUMAN EXPERTS	CHATGPT
• Contextual discrimination in anatomy and judgment	• Broader recall of micro-skills in exposure and hook use
• Stronger for misidentification risk, Rouviere's sulcus, and bailout decisions	• Multicoding mapped co-occurrence, but may overcount without consolidation rules

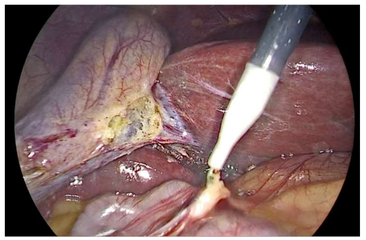
CONCLUSION

• Human experts and ChatGPT were complementary rather than interchangeable.

• Experts added anatomy-sensitive judgment; ChatGPT added scalable pattern recognition and consistent micro-skill detection.

• Challenges include harmonizing operational definitions, managing synonym variability, and regulating multicoding density.





TAKE-HOME MESSAGE

AI-assisted coding may help scale curricular surveillance, but expert oversight remains essential for anatomy, safety, and escalation decisions.