The Quantification of Efficacy and Accuracy for Screening and Data Extraction in Perinatal Mood and Anxiety Disorders Studies: A Comparative Analysis of Traditional vs. GenAl-Powered Approaches

Kyeryoung Lee, PhD 1* , Hunki Paek, PhD 1* , Emma McNeill, MS 2 , Brian Christman, MS 2 , Sabrina Alam, MPH 2 , Lizheng Shi, PhD 3 , Xiaoyan Wang, PhD 1,3 ¹IMO Health, Rosemont, IL, USA, ²University of Mississippi Medical Center, Jackson, MS, USA, ³Tulane University, New Orleans, LA, USA *Equal Contribution



Introduction

Al-Enabled vs. Manual SLR: PMAD Case Study

Objective

Quantify efficacy and accuracy of Al vs. manual SLR in PMAD

Manual SLR (26 articles)

VS

AI-Enabled SLR (Same 26 articles)

Comparison Metrics

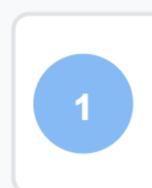
• Time Efficiency • Accuracy • Completeness

Data extraction: study details, model parameters, baseline characteristics, evaluation outcomes

PMAD Study Criteria

Screening/prevention/treatment • Peri/postpartum females • Economic metrics

Methodology



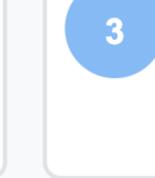
Gold Standard Development

26 manually curated articles as reference standard Expert human review to establish ground truth



Screening Criteria

- · Studies on screening, prevention, or
- treatment for PMAD Peri-or-postpartum females (≤1y)
- Economic evaluation metrics outcomes



Data Extraction Categories

- Study details
- Model parameters
- Baseline characteristics
- Evaluation outcomes (cost, ICER, QALY)



Performance Evaluation

Manual Process Human screening & data extraction Time tracking for process steps Recorded elements extracted

Automated SLR System Al screening & data extraction Time tracking for process steps Recorded elements extracted

Metrics: Time Reduction | Accuracy | Error Correction | F1-Score

Al-Assisted vs. Manual Systematic Review Workflow

Manual Process

Abstract Screening ~100 min/article Human judgment only

Full-Text Review Manual processing Prone to missed articles

Data Abstraction ~300 min/article 20-30 elements captured

Data Accuracy 14 errors detected 36 missed data elements

Al-Assisted Process

Abstract Screening <1 min/article 96.2% accuracy

Full-Text Review Automated processing

100% accuracy

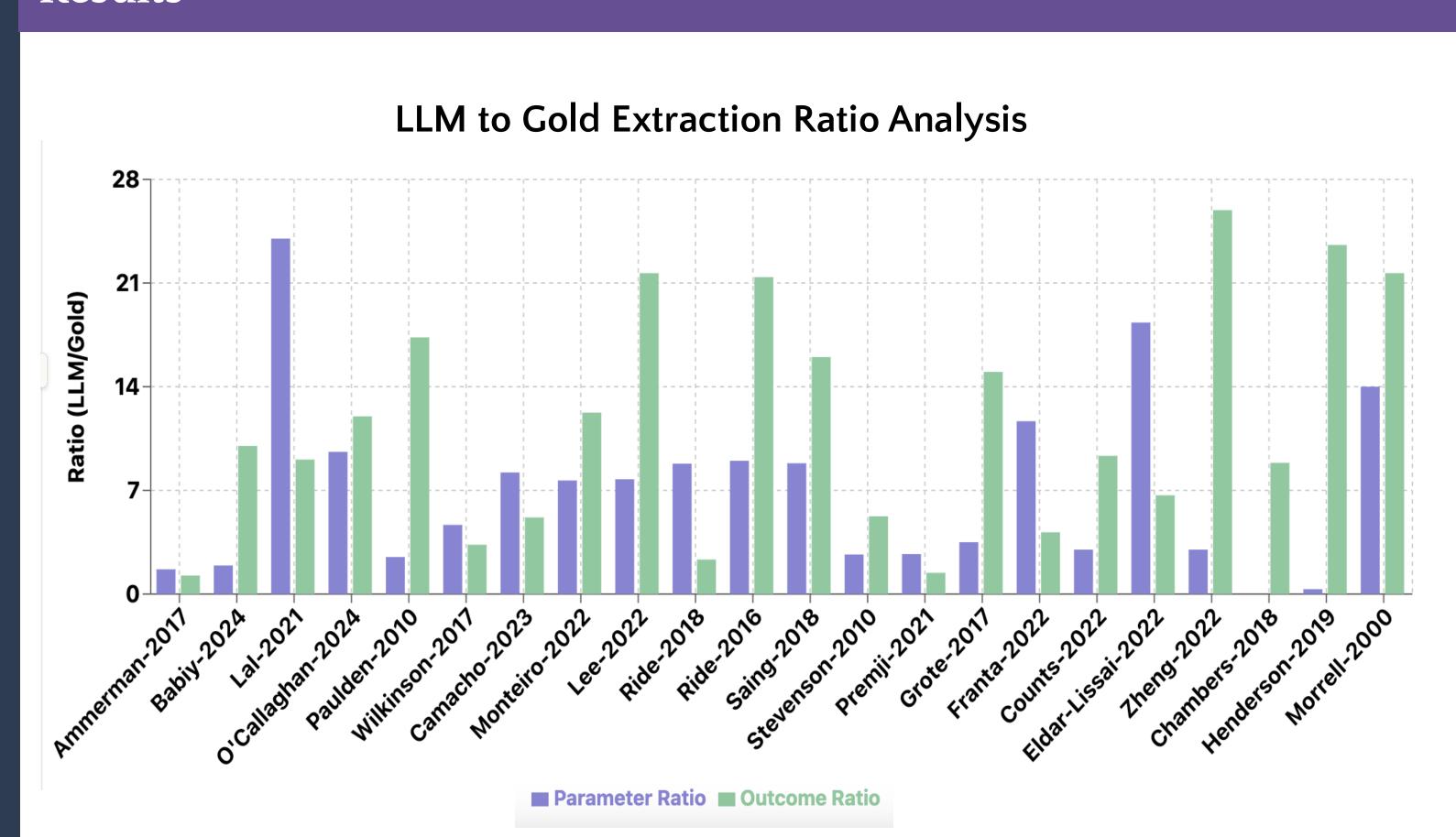
Data Abstraction 10-30 min/article

100-200+ elements captured

Data Accuracy F1-score: 0.993 Error correction capability

Total Time: 400 min \rightarrow <40 min (>90% reduction)

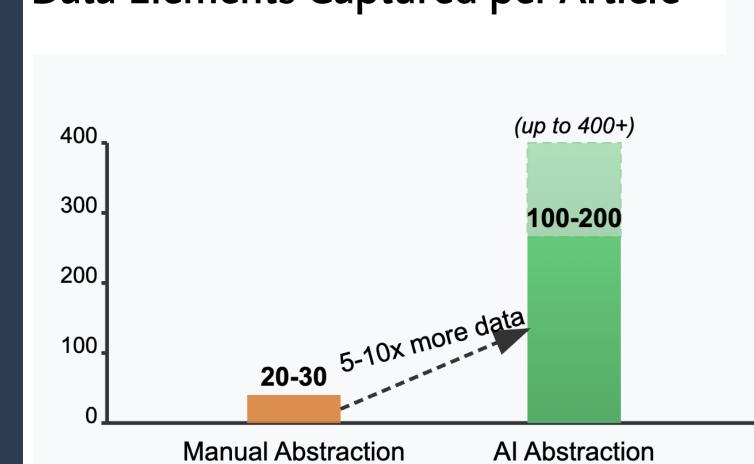
Results



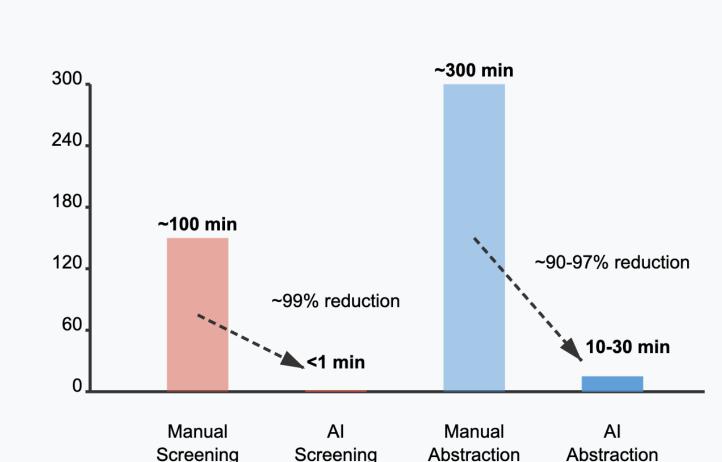
Key Observations:

- •LLM extractions exceed Gold standards by 9.2x for parameters and 10.7x for outcomes
- •In studies with zero Gold parameters, LLM still extracted significant numbers (e.g., Heslin-2022: 72, Petrou-2006: 76; not shown in this ratio graph)
- •Most extreme parameter extraction: Morrell-2009 (7 Gold vs 456 LLM)
- •Most extreme outcome extraction: Zheng-2022 (12 Gold vs 311 LLM)

Data Elements Captured per Article



Time Comparison: Manual vs Al

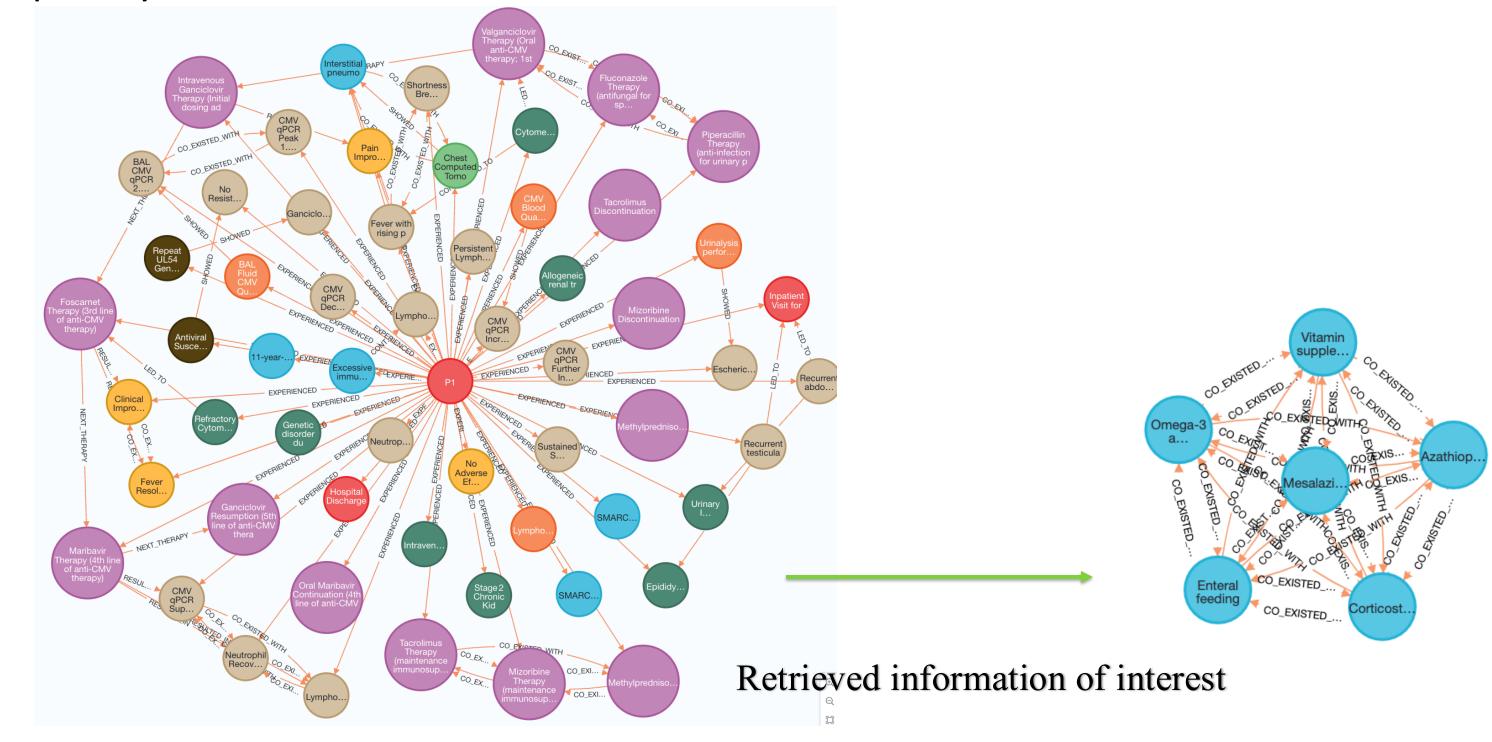


Conclusion

Our AI-system significantly reduced time and error while effectively enhancing accuracy and comprehensiveness in screening and data extraction from full-texts/tables. This approach holds promise to advance SLRs and health economics and outcomes research.

Plan

Construct a knowledge graph using the extracted data for the efficient retrieval and deep analysis



Contact Information

Kyeryoung Lee: <u>klee@imoheal.com</u> hpae@imohealth.com Hunki Paek: Xiaoyan Wang: xwang@imohealth.com

