



EVALUATING LARGE LANGUAGE MODEL-DERIVED PROCEDURE SELECTION AGAINST REAL-WORLD BARIATRIC SURGICAL PRACTICE: A UK NATIONAL REGISTRY ANALYSIS

**Mr Shivam Bhanderi, Dr Paul Ingley, Dr Swarna Kempe Gowda,
Miss Emma Rose McGlone, Professor Omar Khan, Mr Chandra Cheruvu**

North Midlands Upper GI & Bariatric Unit, University Hospitals of North Midlands NHS Trust

National Bariatric Surgery Registry (NBSR)

CONFLICT OF INTEREST



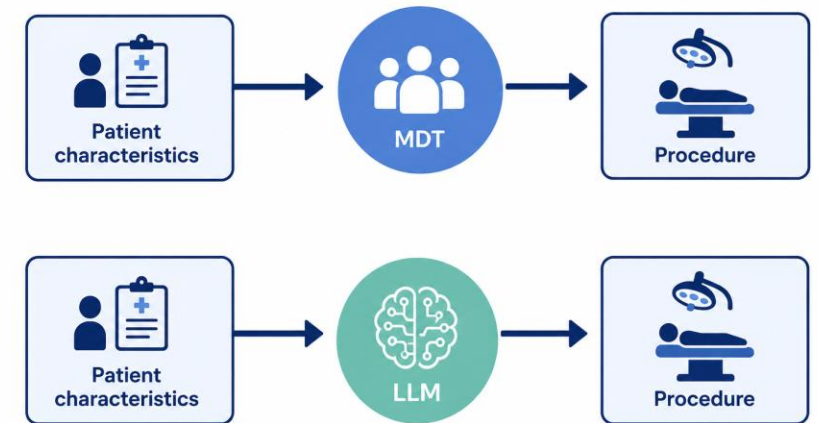
No conflict of interest to declare.

BACKGROUND: WHY ASK THIS QUESTION?

Procedure selection in metabolic bariatric surgery is complex

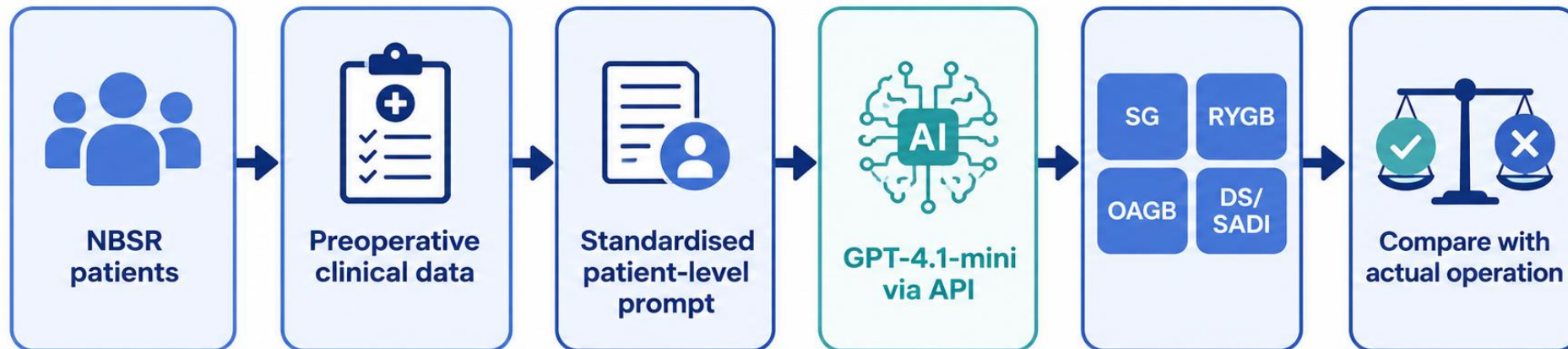
- Integrates patient factors, comorbidities and surgical risk
- Traditionally undertaken through multidisciplinary decision-making
- LLMs are increasingly proposed as clinical decision-support tools

Can an LLM reproduce real-world bariatric procedure selection at scale?

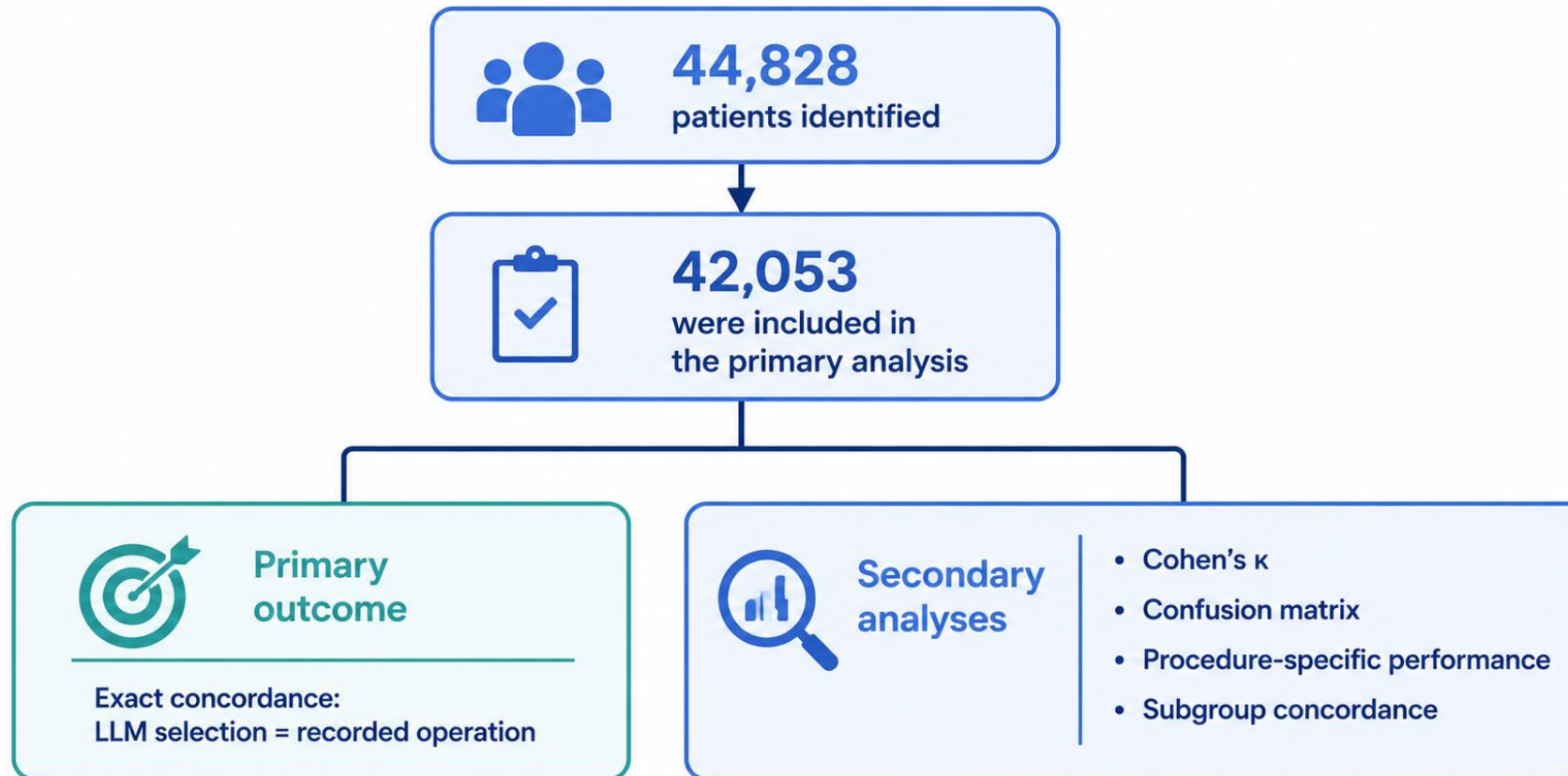


AIM & STUDY DESIGN

To evaluate agreement between LLM-derived procedure selection and real-world bariatric surgical practice using UK national registry data



COHORT & ANALYSIS

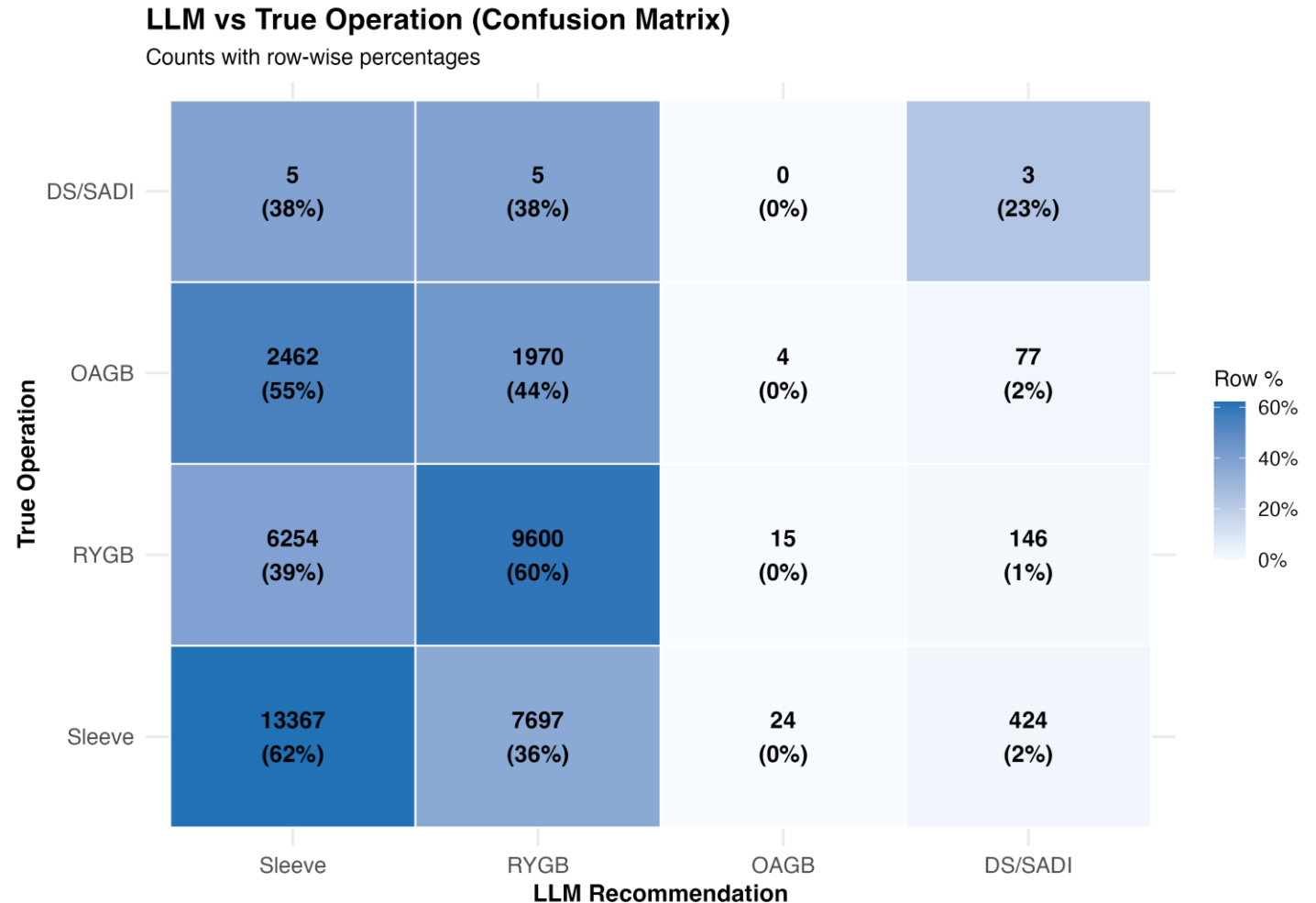


PRIMARY RESULT

- Overall agreement was moderate
- Chance-adjusted agreement was limited
- Overall concordance: 54.6%
- Cohen's kappa $\kappa = 0.185$
- Raw agreement $\neq$ replication of clinical decision-making

Where did the LLM disagree with real-world practice?

- SG ↔ RYGB were frequently confused
- OAGB was rarely recognised
- Systematic rather than random discordance

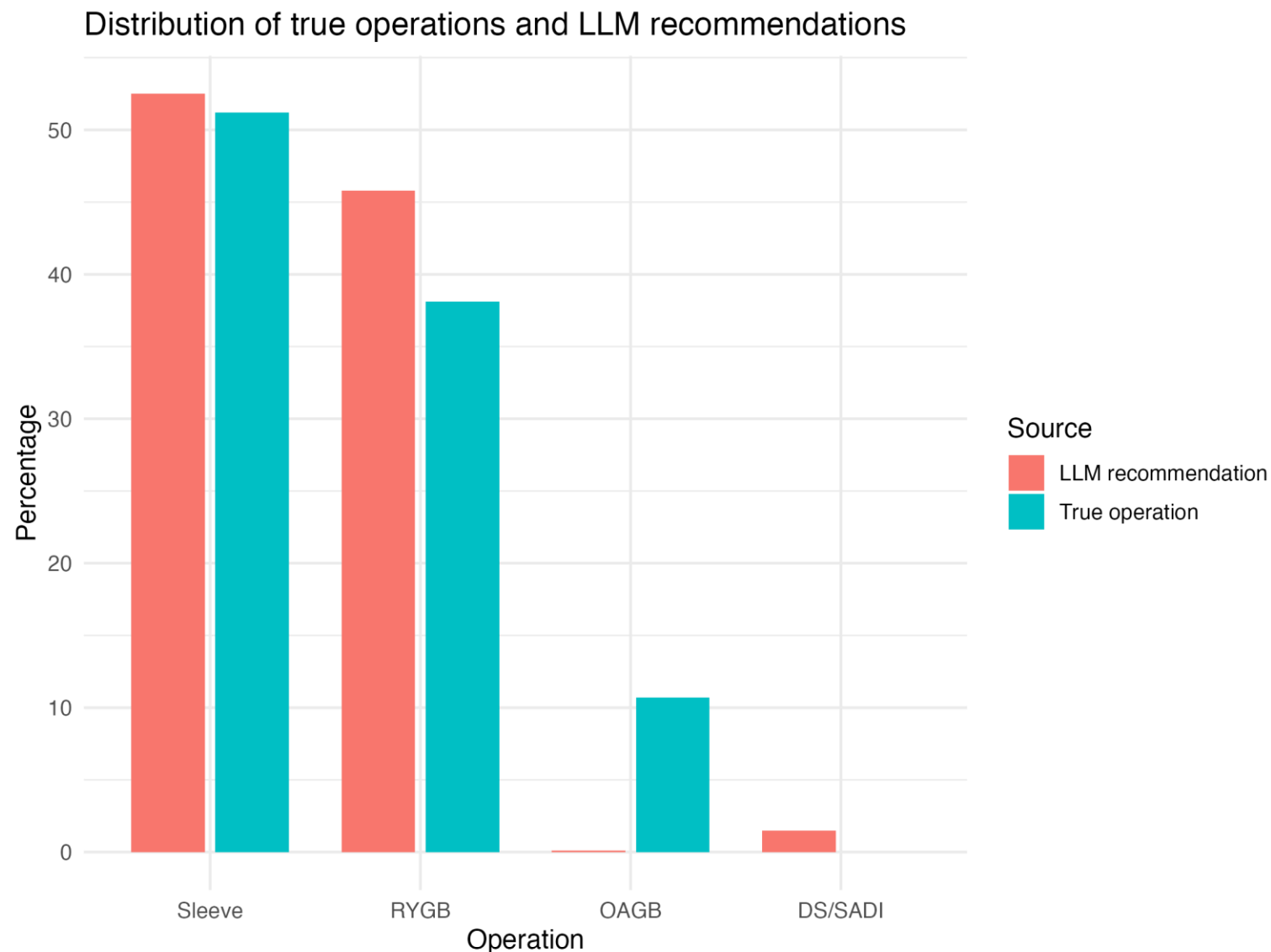


PROCEDURE DISTRIBUTION



LLM recommendations differed systematically from real-world practice

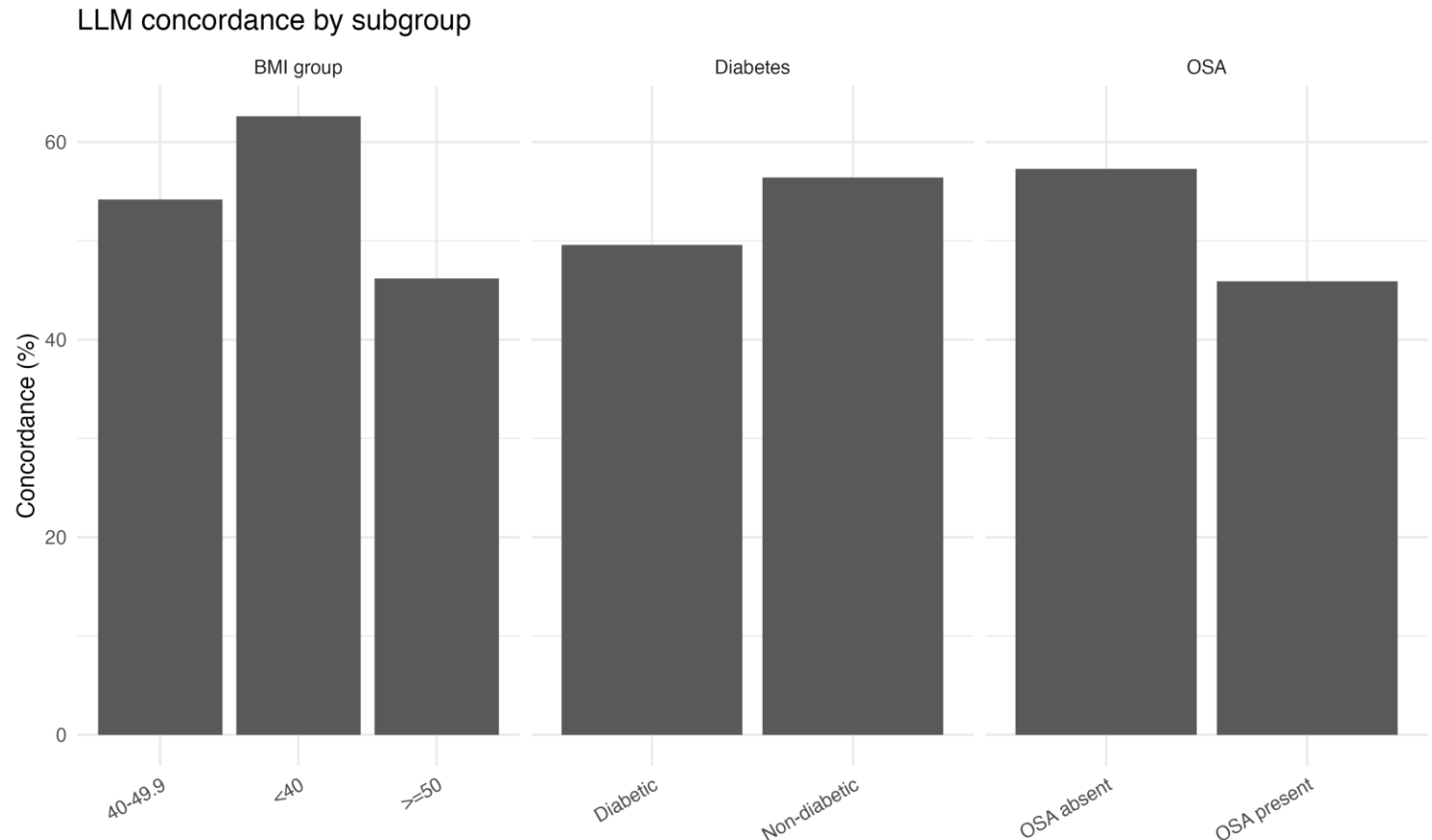
- OAGB: 10.7% actual → 0.1% LLM
- RYGB over-recommended
- DS/SADI over-recommended



SUBGROUP ANALYSIS

Concordance fell with increasing clinical complexity

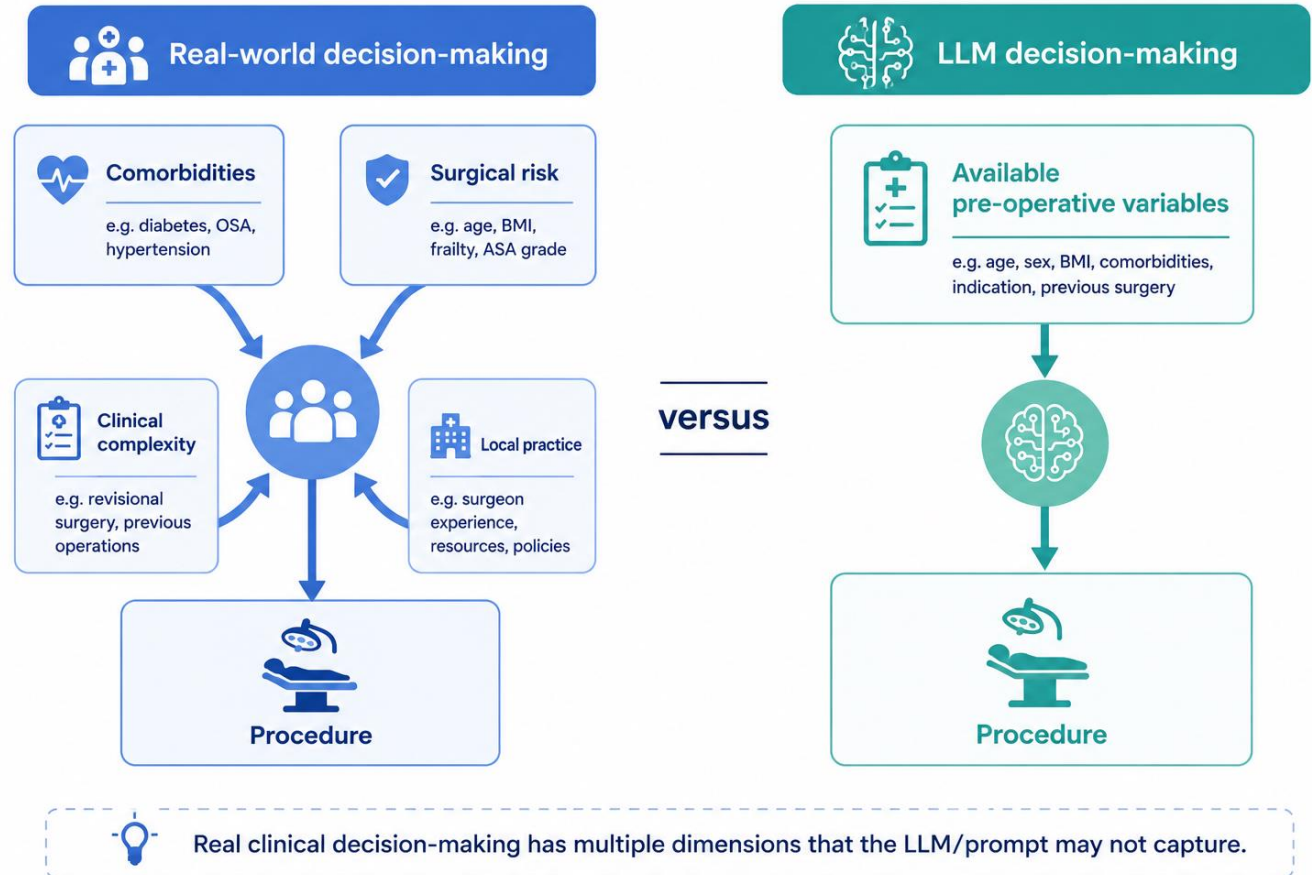
- Higher concordance: Lower BMI / lower clinical complexity
- Lower concordance: Higher BMI / higher ASA
- T2DM influenced selection but did not improve alignment



INTERPRETATION

- 54.6% concordance demonstrates partial alignment with real-world practice
- $\kappa = 0.185$ indicates limited agreement beyond chance
- Systematic procedure-specific bias was evident
- Concordance decreased with increasing clinical complexity

LLM decision-making appears simplified compared with real-world bariatric practice



LIMITATIONS

- Recorded operation used as a **proxy** for real-world procedure selection
- Recorded operation **may differ** from the formal MDT recommendation
- **Centre-specific preferences** and **procedural availability** were not captured
- **Imbalanced procedure prevalence** affects interpretation of concordance and κ
- It does not establish whether either the LLM recommendation or the operation performed was the **clinically optimal choice**

CONCLUSIONS

- LLMs demonstrate partial alignment with real-world bariatric procedure selection
- Systematic biases and limited chance-adjusted agreement remain
- Performance appears poorer in more clinically complex patients

Decision support — not MDT replacement

Further refinement and prospective validation are required before clinical integration

THANK YOU

XXX IFSO WORLD CONGRESS

"Advancing Multimodal Obesity Care"

24-27 August 2027

ifso2027.org



IFSO
PARIS 2027

See you in

PARIS

