Harnessing artificial intelligence in bariatric surgery: comparative

analysis of ChatGPT-4, Bing, and Bard in generating clinician-level

bariatric surgery recommendations

**ChatGPT** is better than Bard and Bing for answering Bariatric surgery questions.

36 clinicians rated each based on Likert score.

(Likert score means 1-5 strongly disagree to strongly agree)

**ChatGPT had 85.7% accuracy** with Likert score >= 4

**ChatGPT had 21.68 FRE readability** (Very difficult to understand; college level syntax)

Bing potentially failed due to search engine accessing incorrect info online.

**Improvements:** Readability (ideally FRE score > 60), accuracy could be better (maybe provide resources to reference)

**Assessment tools:** Excel for data collection, FRE for readability analysis

**Conclusion:** Use ChatGPT for chatbot and improve readability.

Current Applications of Artificial Intelligence in Bariatric Surgery

Covers 36 studies from 2001-2021

Used Area Under Curve (AUC) rating (1 = perfect classification, .5 = random)

Best Options:

Neural Network (NN): AUC = .81 for training but .66 for testing. OVERFIT.

Random Forest (RF): AUC = .92 for training but .52 for testing. OVERFIT.

Predicting difficult intubation: use Xgbc algorithm. Accuracy = 80%, Precision = 100%

Presurgical evaluation:

Obstructive Sleep Apnea (OSA). Polysomnography (PSG) diagnoses OSA.

Chronic Obstructive Pulmonary Disease (COPD), chronic lung disease, asthma. Detected by Spirometry.

Hiatial Hernia (stomach goes up through hiatus into chest). Detect with contrast swallow Swallow Study (SS).

Intraoperative:

Drug concentrations

Remaining surgery duration. Determined by deep learning pipeline RSDNet.

Postoperative:

Page 2724 and 2726

Artificial Intelligence in Bariatric Surgery: Current Status and Future

Perspectives

| **Use Case** | **AUC / Accuracy (Max)** | **Notes** |
| --- | --- | --- |
| Predicting 90-day complications (ANN) | **AUC = 0.98** | Very high, but not externally validated |

Most important variables predicting mortality:

Age, BMI at start, history of heart failure, insulin use, smoking status.

Variables for predicting Hiatial hernias:

Reflux symptoms, higher age, BMI = higher risk

Predicting operation length:

Lower age, lower BMI = shorter operation

Neural Network (NN) best at predicting post operative complications.

The development of machine

learning in bariatric surgery

| **Use Case** | **Model Used** | **Notable Metrics / Observations** |
| --- | --- | --- |
| Predicting postoperative outcomes | ANN, ensemble models | Ensemble models outperformed base models |

Page 3 provides useful methods. Standouts:

ML using Contrast Swallow Studies to predict Hiatial hernia. Improved diagnostic sensitivity by 1.5 times baseline. Preop.

Nudel - ANN good for predicting postoperative gastrointestinal leakage and venous thromboembolism (VTE). High AUC and specificity, but its sensitivity was low

Razzaghi - Ensemble ML, applied with Synthetic Minority Oversampling Technique (SMOTE) has best accuracy for predicting post operative outcomes. It is important to note that

they just considered total and not individual complications (22).

This is due to low sample size and limited complications in their

database

Cao - 3 unsupervised deep learning NN for predicting the occurrence of severe postoperative complications.

CAO - CNN for predicting long term health related life quality after surgery.

Modaresnezhad – RxSem model. used semantic data integration, standardization, and dimensionality reduction that allowed for fast and efficient application of data mining on large clinical datasets.

Zhang - Siamese-KNN (K Nearest Neighbor) ML predicts eventual weight loss 6 months after surgery with 84% accuracy. But uses fMRI images.

Sheidaei- Decision tree good for predicting surgery type based on info from first physical exam.

A machine learning approach to predict types of bariatric surgery using

the patients first physical exam information

Compares Sleeve vs Bypass

Use decision tree

Depth = 5

Accuracy = 77%

Predict 99% of bypass cases (Sensitivity) correctly

Evaluating AI Capabilities in Bariatric Surgery: A Study on ChatGPT‑4

and DALL·E 3’s Recognition and Illustration Accuracy

Note relevant. For image recognition and creation.

ChatGPT tried to interpret surgical illustrations but only got 1/6 correct.

DALLE tried to create surgical illustrations but got 0/6 correct.

International expert consensus on the current status and future prospects of artificial intelligence in metabolic and bariatric surgery

Asked a 68 doctors from 35 countries what they think about AI use in Bariatric Surgery.

Emphasizes the need to design your tool with:

* Clinician oversight
* Explainability (e.g., SHAP values) (SHAP means how much each feature contributed to a particular prediction, and in what direction (positive or negative influence).

Conclusions

**Most Effective AI Models**

| **Model Type** | **Consensus Performance** | **Strengths** | **Weaknesses** |
| --- | --- | --- | --- |
| **Neural Networks (ANN/CNN)** | High in some tasks | Captures nonlinear patterns, adaptable | Requires tuning, overfitting risk |
| **Random Forest** | Strong baseline | Interpretable, ensemble robustness | Can lag in nuance for complex data |
| **Gradient Boosting (GBM/XGBoost)** | High accuracy | Good with structured data, low bias | Harder to interpret, prone to overfitting |
| **Logistic Regression (LR)** | BAD | Simple, interpretable | Weak with nonlinear or noisy features |
| **Support Vector Machines (SVM)** | BAD | Works with smaller datasets | Slower, brittle with large data |

*Conclusion*: **ANNs and ensemble methods** (like Random Forest and GBM) were most successful for outcome prediction and risk assessment in bariatric surgery.

**Limitations**

* **Retrospective datasets** (5/6 papers) → prone to bias. (Bias such as inaccurate/missing data, hidden variables such as smoking)
* **Low interpretability** (aka black box) for many NN-based solutions. (Fix with SHAP)

| **Recommendation** | **Rationale** |
| --- | --- |
| Start with Random Forest or XGBoost | Easier to tune, high accuracy, and widely accepted in healthcare |
| Use ANN for more complex feature sets | For deep EHR data or sequential data like time-series vitals |

**Workflow**

1. **Data Cleaning**: Handle missing values, encode categories.
2. **Feature Engineering**: Include vitals, labs, demographics, comorbidities.
3. **Split Data**: Train/test/validation with stratification meaning distributing classes evenly amongst datasets
4. **Model Training**:
   * Start with **Random Forest / XGBoost**
   * Then prototype **ANN** for more complexity
5. **Explainability**: Use **SHAP** to interpret black-box models.
6. **Validation**: If possible, use a second dataset or simulate cross-hospital testing.