Value-Based Healthcare (VBHC) in Breast Reconstruction Surgery: Leveraging LLM to understand cost context from PubMed literature

Suryanarayana Jayanthi, Alphonse Chander
The University of Texas at Austin

ABSTRACT

Breast reconstruction after mastectomy is a critical aspect of patient care, significantly affecting the psychological well-being of patients [15, 16]. However, the complex landscape of surgical techniques, outcomes, and cost-effectiveness data in the breast reconstruction literature poses a substantial challenge for clinicians seeking optimal treatment strategies [2, 3]. This project introduces a novel Retrieval-Augmented Generation (RAG) pipeline, using GPT-4, to automate and enhance the review of the literature in breast reconstruction. Our system effectively retrieves and synthesizes evidence from PubMed articles, providing structured clinical insights on outcomes, quality-adjusted life years (QALY), and patient satisfaction [1]. The RAG pipeline employs vector embeddings for efficient information retrieval and GPT-4 for generating coherent and contextually relevant summaries [9, 10]. This innovative approach aims to support evidence-informed value-based surgical decision making, addressing clinical questions that are difficult to resolve through traditional manual review, with the potential to transform clinical workflows and improve patient outcomes.

KEYWORDS

RAG, GPT-4, Breast Reconstruction, Literature Summarization, FAISS, Clinical NLP, Cost-Effectiveness, Evidence-Based Medicine

REFERENCES TO OTHER DOCUMENTS FOR SUBMISSION

Below list gives links to urls

- CODE
- PRESENTATION
- VIDEO

1 INTRODUCTION

Breast cancer is a prevalent malignancy, often requiring mastectomy, surgical removal of the breast. Breast reconstruction, a surgical procedure to restore the breast's appearance following mastectomy, plays a crucial role in the physical and psychological recovery of patients. It contributes to improved self-image, body confidence, and overall emotional well-being. Numerous studies have consistently demonstrated the positive impact of breast reconstruction on patients' quality of life, particularly in alleviating the psychological distress associated with breast cancer diagnosis and treatment.

Selecting the most appropriate method for breast reconstruction is a complex and highly individualized decision-making process. Techniques for reconstruction vary significantly in approach, invasiveness, and outcomes, and can be broadly classified into three categories:

1.1 Implant-Based Reconstruction

This method involves the placement of saline or silicone implants to recreate the breast mound. It is less invasive and generally entails shorter operative and recovery times. However, complications such as capsular contracture, implant rupture, infection, and issues with symmetry are not uncommon, particularly in patients who undergo radiation therapy.

1.2 Autologous Tissue Reconstruction

Autologous techniques utilize the patient's own tissue—typically harvested from the abdomen (e.g., DIEP, TRAM flaps), back, or thighs—to construct a more natural-looking breast. While autologous procedures are associated with longer operative times and increased perioperative complexity, they generally offer superior aesthetic outcomes and higher patient satisfaction over the long term.

1.3 Hybrid Approaches

Hybrid reconstruction combines elements of both implant-based and autologous methods to balance the benefits and limitations of each. These approaches may be employed to enhance cosmetic outcomes, address anatomical constraints, or optimize volume replacement following partial mastectomy.

The optimal reconstructive strategy depends on multiple clinical and personal variables, including patient anatomy, prior oncologic treatment, comorbidities, lifestyle, and aesthetic goals. In this multifaceted decision landscape, evidence-based planning is essential.

Traditional methods of evidence synthesis—such as manual reviews, consensus guidelines, and meta-analyses—struggle to keep pace with the exponential growth of biomedical publications. These approaches are often time-consuming, inflexible, and do not provide real-time insights tailored to specific clinical questions. As a result, clinicians and healthcare administrators frequently lack immediate access to comparative effectiveness data that could inform nuanced surgical decisions, potentially impacting both clinical outcomes and healthcare costs.

To address this challenge, we introduce an AI-powered literature question answer system based on Retrieval-Augmented Generation (RAG). Our system processes full-text biomedical PDFs from pubmed literature and generates concise, evidence-linked responses to complex clinical questions, with a focus on key surgical outcome domains: cost-effectiveness, complication rates, and patient-reported outcomes. By accelerating and scaling evidence synthesis, this approach aims to support informed, patient-centered decision-making in breast reconstruction.

2 RELATED WORK

Natural language processing (NLP) in the biomedical domain has evolved rapidly in recent years, driven by advances in transformer-based models and the increasing availability of large-scale medical corpora. Pre-trained language models such as BioBERT [4], ClinicalBERT [5], and BioGPT [6] have demonstrated superior performance across a range of NLP tasks, including named entity recognition, relation extraction, and biomedical question answering. These models are typically trained on domain-specific text such as PubMed abstracts and clinical notes, enabling them to better understand medical terminology and context.

In addition to representation learning, biomedical question answering systems such as BioASQ [8] and PubMedQA [7] have aimed to extract or generate answers to natural language queries using structured benchmarks. However, these systems often operate at the abstract level and do not incorporate full-text processing, limiting their utility in complex clinical decision-making. They typically rely on extractive approaches and lack the ability to synthesize cross-document evidence or provide structured insights into nuanced clinical topics.

Retrieval-Augmented Generation (RAG) [9] has recently emerged as a powerful technique that integrates dense vector retrieval (e.g., via FAISS [10]) with generative language modeling. In RAG architectures, a user query is used to retrieve semantically similar document segments from a corpus, which are then used to condition the generation of answers by models such as GPT-3 or GPT-4. This technique has been shown to improve factual accuracy and reduce hallucinations, especially in knowledge-intensive domains like medicine. Our system builds on this paradigm by applying RAG to the full-text breast reconstruction literature, enabling contextualized, evidence-backed answers to clinical questions.

In parallel, extensive research has been conducted in the field of breast reconstruction, focusing on comparative effectiveness, patient-reported outcomes, and cost-efficiency. Multiple studies have evaluated differences in clinical outcomes between implant-based and autologous reconstruction methods, analyzing complication rates, surgical revisions, and long-term aesthetic results [2]. Tools such as the BREAST-Q have become standard for assessing patient-reported outcome measures (PROMs), offering a structured way to evaluate patient satisfaction and quality of life. Cost-effectiveness analyses, including QALY-based models, have been used to compare the economic value of DIEP and TRAM flaps [1, 3], guiding healthcare policy and reimbursement strategies.

Despite the richness of the existing literature, synthesizing insights across studies remains labor-intensive and slow. There is a growing need for intelligent systems that can bridge the gap between evidence and decision-making. Our work addresses this need by automating literature synthesis using a RAG framework tailored to reconstructive surgery outcomes, allowing dynamic access to high-quality, contextualized evidence in real time.

3 METHODOLOGY

We designed and implemented a modular Retrieval-Augmented Generation (RAG) pipeline specifically tailored for the task of synthesizing clinical insights from breast reconstruction literature. The system combines dense vector retrieval using FAISS with large language model-based summarization via GPT-4. Below, we describe each step in detail:

3.1 Document Ingestion

We curated a dataset of 12 peer-reviewed, full-text scientific articles from PubMed, focusing on comparative studies in breast reconstruction. These documents were parsed using the PyPDF2 library, which allowed extraction of continuous, structured text from PDF files.

3.2 Text Chunking

To facilitate semantic retrieval, each article was divided into overlapping chunks of approximately 1000 tokens with a stride of 200 tokens. This strategy ensured that important context, especially for clinical outcomes and cost data, was preserved across boundaries. Chunking was performed using the 'RecursiveCharacterTextSplitter' utility from LangChain, optimizing for both readability and embedding performance.

3.3 Vector Embedding

Each chunk was transformed into a 1536-dimensional vector using OpenAI's 'text-embedding-ada-002' model. These embeddings capture the semantic content of each segment and enable comparison with clinical queries in a shared vector space. The use of OpenAI embeddings was chosen for their proven performance in zero-shot biomedical retrieval tasks and compatibility with LangChain infrastructure.

3.4 Semantic Indexing with FAISS

The vectorized document chunks were indexed using Facebook AI Similarity Search (FAISS), which supports approximate nearest neighbor retrieval at scale. FAISS was configured with an L2 distance metric and 'IndexFlatL2' backend to ensure fast and accurate semantic search. This index served as the knowledge base for all downstream query resolution tasks.

3.5 Retrieval-Augmented Question Answering

User queries were framed as natural language clinical questions, such as "What complications are associated with increased length of stay after DIEP flap reconstruction?" These queries were passed into a LangChain 'RetrievalQA' pipeline, which first retrieved the top-k most relevant document chunks from the FAISS index. These retrieved passages were then supplied as context to the GPT-4 model via OpenAI's chat-completion API. Prompt engineering was employed to ensure that GPT-4 responses included structured outputs, such as cost data, complication profiles, and citations where applicable.

3.6 Workflow Overview

Table 1 outlines the end-to-end process, from document ingestion to final answer synthesis. The pipeline was designed for modularity, allowing future adaptation to other surgical domains or question formats.

Table 1: RAG-Based Clinical Literature Workflow

Stage	System Operation	Component
Upload full-text PDFs	Extract text and remove artifacts	PyPDF2
Chunk articles	Generate overlapping 200-token segments	LangChain TextSplitter
Embed text chunks	Create semantic vectors using OpenAI API	‘text-embedding-ada-002‘
Index embeddings	Store in FAISS for fast similarity search	FAISS (IndexFlatL2)
Input clinical query	Retrieve relevant document chunks	LangChain Retriever
Generate structured response	Use GPT-4 to summarize findings	OpenAI Chat API

3.7 Evaluation Strategy

To assess the quality of generated responses, we implemented a two-phase evaluation process:

- **Factual Alignment:** Each answer was manually cross referenced with its source chunk to verify fidelity and avoid hallucination.
- **Clinical Usefulness:** Responses were informally reviewed by the research team for clarity, completeness, and relevance to the clinical questions. Answers were compared qualitatively against expected clinical reasoning patterns.

The pipeline was also tested for reproducibility by repeating queries across sessions and comparing response variance.

4 RESULTS

The RAG-based pipeline successfully answered a diverse range of clinical questions by synthesizing evidence from full-text articles and structuring responses around surgical outcomes, cost-effectiveness, and patient-reported measures. Below, we present representative examples of questions and the system's corresponding output, demonstrating its ability to deliver nuanced, evidence-based summaries grounded in peer-reviewed literature.

4.1 Case 1: Postoperative Complications and Length of Stay

"What complications are associated with increasing length of stay after microvascular breast reconstruction?"

The pipeline retrieved and analyzed data from multiple sources related to inpatient recovery, costs, and complication rates. GPT-4 synthesized the following answer:

> “The context does not directly state the complications associated with an increased length of stay after microvascular breast reconstruction. However, it does mention several risk factors, including obesity, diabetes, malignancy history, operative time, a history of radiation therapy, smoking, and bilateral reconstruction, that may necessitate a longer length of stay. These risk factors could potentially lead to complications, but the text does not specify what those complications might be.” [1, 2]

This response demonstrates GPT-4's ability to reason cautiously based on available evidence, avoiding unsupported inferences while still identifying relevant risk factors.

4.2 Case 2: DIEP vs TRAM Flap Cost-Effectiveness

"Are DIEP flaps more cost-effective than TRAM flaps?"

The system synthesized findings from cost-utility analyses comparing long-term outcomes of DIEP and TRAM flap techniques:

> “The prevailing notion that TRAM flaps are more cost-effective than DIEP flaps seems to be contradicted by the study's results. The total hospital charges to costs using cost-to-charge ratio and hospital stays associated with TRAM and DIEP flaps were found to be similar. Specifically, the total hospital charges to costs were $29,775 for DIEP and $28,466 for TRAM. Therefore, the study suggests that the costs between TRAM and DIEP flap surgeries are generally equal. However, it is important to note that factors such as income quartile, primary payer of hospitalization, and geographic region significantly influenced flap choice. This data suggests that socioeconomic factors also play a role in the cost-effectiveness of these procedures.” [3]

This answer goes beyond the traditional clinical lens and highlights the influence of systemic variables—including insurance status and geography—on what is often perceived as a purely clinical cost-effectiveness decision.

4.3 Case 3: PROMs and Technique Selection

"What reconstruction techniques result in higher patient satisfaction scores on the BREAST-Q [11]?"

The pipeline retrieved PROM-centric studies and produced this summary:

> “Reconstruction techniques that result in higher patient satisfaction scores on the BREAST-Q [11] include autologous reconstruction, TRAM reconstruction, immediate LD, pedicled and free TRAM, DIEP, and SIEA flaps. All of these techniques were reported to have statistically significant increases in patient satisfaction when compared to implant-based reconstruction. This was supported by studies conducted by Santosa et al. [15], Hu et al. [16], and Jeevan et al. [17]. These findings were consistent among patients who underwent either immediate or delayed reconstruction. It's important to note that these results are based on the

context provided and may not include all possible reconstruction techniques." [2]

The model highlights a broad consensus across studies that autologous techniques—including DIEP, TRAM (pedicled and free), SIEA, and LD flaps—are associated with significantly higher BREAST-Q satisfaction scores than implant-based reconstruction, regardless of reconstruction timing.

4.4 Cross-Query Insights and Observations

Across over 15 clinical queries tested, the system demonstrated high consistency in:

- Extracting statistically grounded insights from the retrieved documents.
- Presenting trade-offs between techniques, including cost, complications, and quality-of-life implications.

In informal evaluation, 13 out of 15 responses were deemed "clinically useful" and "informative," particularly for planning and comparative analysis. Remaining issues were related to either overly generalized phrasing or incomplete referencing in edge cases where document retrieval failed to retrieve a relevant passage.

4.5 User Experience and Interpretability

Each query resulted in responses generated within seconds, including citations and structured summaries. This format was noted as particularly helpful for clinicians who require synthesized evidence at the point of care. The model's transparency—through citation links and source context—added trustworthiness to its recommendations, although domain oversight remains essential before clinical application.

5 DISCUSSION

The results of this study highlight the practical potential of combining dense vector retrieval with generative large language models for clinical literature synthesis. The system exhibited particularly strong performance on comparative clinical questions that required reasoning across multiple outcome dimensions—such as postoperative complications, patient-reported outcomes, and cost-effectiveness. This multifaceted integration is rarely supported by traditional keyword-based search engines or systematic review tools.

5.1 Synthesis Across Diverse Domains

A key advantage of the RAG pipeline is its ability to synthesize evidence from heterogeneous sources and present it in a unified, structured form. For instance, the system could contextualize surgical decisions by correlating BREAST-Q subscale scores with complication rates and cost per quality-adjusted life year (QALY). Such synthesis would typically require hours of manual review and interpretation but was performed in near real time using our framework. This functionality has the potential to assist both frontline surgeons and administrative stakeholders engaged in value-based care planning.

5.2 Interpretability and Trust

The system's responses were grounded in context retrieved from semantically similar document chunks, improving the transparency of the underlying evidence base. While the outputs did not include explicit inline citations, answers reflected themes and content from relevant source materials, enhancing interpretability. By limiting the model's generation scope to retrieved context, the pipeline reduced the risk of hallucinated or unsupported claims—a common issue in large language models. Nonetheless, human oversight remains essential, particularly when applying outputs to high-stakes clinical scenarios.

5.3 Limitations and Technical Constraints

Despite its strengths, the system has several limitations:

- **PDF Extraction Constraints:** The pipeline currently relies on clean, machine-readable PDF text. Scanned documents, complex figure layouts, or multi-column formats may reduce extraction fidelity. Integrating Optical Character Recognition (OCR) and layout-aware parsers could address this limitation.
- **Generative Hallucinations:** Although grounded in retrieved content, GPT-4 occasionally generated extrapolations or oversimplified conclusions not explicitly present in the source. Prompt engineering and reinforcement from expert feedback may help mitigate this issue.
- **Domain Adaptation:** The system was not fine-tuned on clinical question-answering datasets. As a result, domain-specific nuances (e.g., temporal follow-up windows, surgical sub-cohorts) may not always be captured accurately.

5.4 Future Directions

Improving the factual robustness of generated responses remains a key priority. Future iterations could benefit from:

- Fine-tuning on clinically annotated QA datasets (e.g., MedQA, HealthQA)
- Implementing citation-aware generation to verify each sentence's source
- Using multi-hop retrieval for questions that require synthesizing across multiple documents
- Adding explainability layers that show which retrieved chunk supports which part of the answer

Ultimately, this study demonstrates that RAG-based systems can play an instrumental role in accelerating literature synthesis, particularly in specialties like reconstructive surgery where outcomes are multi-dimensional and evolving. With appropriate oversight, such systems could reduce evidence bottlenecks and empower more informed, patient-centered decisions at the point of care.

6 CONCLUSION AND FUTURE WORK

This project presents a novel Retrieval-Augmented Generation (RAG) pipeline that integrates state-of-the-art natural language processing with semantic information retrieval to support evidence synthesis in surgical decision-making. By automating the extraction of clinically relevant insights from full-text biomedical literature, the system reduces the cognitive and temporal burden associated with manual reviews. In doing so, it facilitates access to structured, patient-centered information related to surgical outcomes, costs, and satisfaction measures—paving the way for more data-informed and value-based care in breast reconstruction.

Looking forward, we envision several key areas for expansion and refinement:

- **Scaling to Larger Literature Bases:** The current system was developed and evaluated on a curated set of 12 articles. Future iterations will incorporate automated PubMed querying and ingestion to support ongoing updates and enable scaling to hundreds or thousands of relevant publications. This will allow the system to maintain currency with emerging studies and meta-analyses.
- **Real-Time Clinical Integration:** To enhance accessibility and usability, we aim to develop a user-friendly, web-based interface that allows clinicians to interact with the RAG system via natural language queries. This interface will prioritize usability in surgical clinics and tumor boards, offering features such as source tracing, query history, and citation export.
- **Domain Expansion:** While the current focus is on core surgical reconstruction questions, future versions will incorporate adjacent clinical domains that significantly influence decision-making. These include topics such as radiation therapy timing [12], use of acellular dermal matrices (ADM) [13], and the trade-offs between immediate and delayed reconstruction strategies [14].

By continuing to evolve this system toward clinical utility and interdisciplinary integration, we aim to contribute meaningfully to the development of intelligent tools that bridge the gap between medical literature and practice, improving the precision and timeliness of surgical care.

ACKNOWLEDGMENTS

The authors would like to thank Professor Dr. Ying Ding, Hung Le, and the entire teaching assistant team for their valuable guidance and support throughout the course of this project.

CONFLICTS OF INTEREST

The authors affirm that there are no competing financial interests or personal relationships that could have been perceived to influence the research or its findings.

This statement ensures transparency and affirms that the research findings are unbiased and free from external conflicts of interest.

REFERENCES

[1] Alexander F. Mericli et al. The Optimal Length of Stay after Microvascular Breast Reconstruction: A Cost-Utility Analysis. *Plastic and Reconstructive Surgery*, 150(2):279e–287e, 2022. 10.1097/PRS.0000000000009316.

[2] Evan Matros et al. Cost-Effectiveness Analysis of Implants versus Autologous Perforator Flaps Using the BREAST-Q. *Plastic and Reconstructive Surgery*, 135(4):937–945, 2015. 10.1097/PRS.0000000000001134.

[3] Jose A. Foppiani et al. The value of microvascular breast reconstruction: Cost equivalence of TRAM and DIEP flaps. *Microsurgery*, 44(4):e31185, 2024. 10.1002/micr.31185.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 10.1093/bioinformatics/btz682.

[5] Emily Alsentzer, John R. Murphy, William Boag, et al. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78, 2019.

[6] Renqian Luo, Xiaodong Liu, Yichong Xu, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Nature Communications*, 13, 7276, 2022. 10.1038/s41467-022-34875-1.

[7] Qiao Jin, Bhanu Pratap Singh Rawat, and Hong Yu. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2567–2577, 2019.

[8] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, et al. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, 2015.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 9459–9474, 2020.

[10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 10.1109/TBDATA.2019.2921572.

[11] Robert Phan, David J. Hunter-Smith, Warren M. Rozen. The use of Patient Reported Outcome Measures in assessing patient outcomes when comparing autologous to alloplastic breast reconstruction: a systematic review. *Gland Surgery*, 8(4):452–460, 2019. 10.21037/gs.2019.07.04.

[12] Johanna H. Yun, Roberto Diaz, and Amber G. Orman. Breast Reconstruction and Radiation Therapy. *Cancer Control*, 25:1–7, 2018. 10.1177/1073274818795489.

[13] Naveen M. Krishnan, Abhishek Chatterjee, Kari M. Rosenkranz, et al. The cost effectiveness of acellular dermal matrix in expander-implant immediate breast reconstruction. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 67:468–476, 2014. 10.1016/j.bjps.2013.12.035.

[14] Ahmad Makki, Jørn B. Thomsen, Gudjon L. Gunnarsson, et al. A cost-effectiveness analysis of delayed breast reconstruction with pedicled flaps from the back. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 75:2211–2218, 2022. 10.1016/j.bjps.2022.02.034.

[15] Kavita B. Santosa, Ji Qi, Hyungjin M. Kim, et al. Long-term patient-reported outcomes in postmastectomy breast reconstruction: an analysis using the BREAST-Q. *JAMA Surgery*, 153(10):891–899, 2018. 10.1001/jamasurg.2018.1677.

[16] Elizabeth S. Hu, Peter G. Pusic, Andrea L. Waljee, et al. Patient-reported outcomes in breast reconstruction: a systematic review of the literature. *Journal of the American College of Surgeons*, 211(5):767–782, 2010. 10.1016/j.jamcollsurg.2010.06.019.

[17] Ranjeet Jeevan, Joanna P. Browne, Christine Gulliver-Clarke, et al. Surgical determinants of patient-reported outcomes following postmastectomy reconstruction in women with breast cancer. *Plastic and Reconstructive Surgery*, 139(5):1036e–1045e, 2017. 10.1097/PRS.0000000000003258.