Bala *et al. Art Int Surg.* 2025;5:476-89
DOI: 10.20517/ais.2025.44

Artificial Intelligence Surgery

Original Article

Open Access

Clinical insight from mesh implant narratives using zero-shot Retrieval-Augmented Generation approach

Indu Bala[1], Ekta Sharma[2], Lewis Mitchell[1]

[1]School of Computer and Mathematical Sciences, The University of Adelaide, Adelaide 5005, Australia.
[2]School of Mathematics, Physics, and Computing, University of Southern Queensland, Springfield 4300, Australia.

Correspondence to: Dr. Indu Bala, School of Computer and Mathematical Sciences, The University of Adelaide, Adelaide 5005, Australia. E-mail: iinduyadav@gmail.com; indu.bala@adelaide.edu.au

How to cite this article: Bala I, Sharma E, Mitchell L. Clinical insight from mesh implant narratives using zero-shot Retrieval-Augmented Generation approach. *Art Int Surg.* 2025;5:476-89. https://dx.doi.org/10.20517/ais.2025.44

Received: 8 May 2025 **First Decision:** 14 Jul 2025 **Revised:** 20 Aug 2025 **Accepted:** 11 Sep 2025 **Published:** 31 Oct 2025

Academic Editor: Gaya Spolverato **Copy Editor:** Xing-Yue Zhang **Production Editor:** Xing-Yue Zhang

Abstract

Aim: Mesh implant surgeries for hernia repair are frequently associated with adverse events that can compromise patient outcomes. Extracting structured clinical insights from large-scale, unstructured data sources such as the U.S. Food and Drug Administration's Manufacturer and User Facility Device Experience (FDA MAUDE) database remains a challenge due to variability and subjectivity in patient narratives. This study aims to develop and evaluate a zero-shot generative artificial intelligence (AI) framework enhanced with Retrieval-Augmented Generation (RAG) to automatically extract structured clinical information and adverse event indicators from unstructured mesh implant reports, assessing its accuracy, interpretability, and scalability against a manually annotated benchmark.

Methods: The study employed the LLaMA 2 (13B) model for zero-shot structured summarization and adverse event extraction from FDA MAUDE mesh implant reports (2000–2021). The framework integrated retrieval-based context using RAG and evaluated model performance on report date, hernia type, and adverse event flag using accuracy, Jaccard similarity, and Chi-square tests ($P < 0.05$). Statistical analysis validated improvements in output reliability and clinical relevance.

Results: The model outputs were compared to a manually annotated Benchmark Baseline. With zero-shot prompting alone, the model achieved accuracies of 67% for report date, 60% for hernia type, and 83% for adverse event flag. After integrating the RAG approach, these accuracies improved to 81%, 82%, and 99%, respectively. The accuracy for adverse event extraction increased from 60% to 86%, and the Jaccard similarity improved from

75% to 88.9%. Chi-square tests confirmed statistical significance ($P < 0.05$) for most of the observed improvements.

Conclusion: This study demonstrates that combining zero-shot generative AI with retrieval augmentation can effectively convert unstructured patient reports into structured data. This approach offers a scalable and interpretable method for adverse event monitoring in mesh implant surgeries and supports data-driven evaluation of patient-reported outcomes.

Keywords: Mesh implantation, adverse events, Retrieval-Augmented Generation (RAG), hernia repair, GenAI

INTRODUCTION

Hernia mesh implants have become a cornerstone of modern surgical practice for repairing inguinal, umbilical, incisional, and stress urinary incontinence (SUI)-related hernias. These devices are designed to reinforce weakened tissue and reduce recurrence rates, thereby improving surgical outcomes and patient recovery[1]. However, despite their benefits, mesh implants are frequently associated with postoperative complications such as chronic pain, infection, erosion, migration, bowel obstruction, and urinary dysfunction[2]. These adverse events can severely affect patients' quality of life, increase healthcare costs, and raise long-term safety concerns. As awareness of these risks grows, there has been an increased focus on post-market surveillance systems to capture, analyze, and respond to real-world patient experiences in a timely manner.

The U.S. Food and Drug Administration's (FDA) Manufacturer and User Facility Device Experience (MAUDE) database provides an invaluable resource for this purpose[3]. It contains thousands of patient-generated narratives detailing mesh-related complications, offering granular insights into device performance in clinical practice. However, these reports remain underexploited for regulatory science and safety monitoring. The primary challenge lies in their format: narratives are unstructured free-text with inconsistent terminology, varying linguistic styles, and significant subjectivity[4]. Extracting clinically meaningful, structured insights from such data is challenging but essential to support evidence-based decision making for both regulators and clinicians.

Traditional natural language processing (NLP) approaches have attempted to address this challenge, typically using rule-based extraction systems, keyword searches, or supervised machine learning models trained on annotated corpora[5-7]. However, these methods face persistent limitations in handling linguistic variability, negation, synonymy, and implicit mentions of adverse events[8]. Additionally, they depend heavily on expensive, domain-specific labeled datasets, which restrict their scalability and adaptability across different medical contexts[9]. As a result, adverse events that are described implicitly or in non-standard terminology may remain undetected, leading to underestimation of device-related complications in surveillance data.

In recent years, large language models (LLMs) such as Meta's LLaMA 2 have shown remarkable promise in addressing these limitations. With their ability to perform zero-shot prompting, LLMs can execute tasks based solely on natural language instructions, eliminating the dependence on large annotated datasets[10-13]. This is particularly advantageous in domains such as mesh implant surveillance, where specialized labeled corpora are scarce or unavailable[14]. However, despite these advantages, LLMs are prone to factual inconsistency and hallucinations when operating without external knowledge grounding, posing significant risks in safety-critical domains where accuracy and interpretability are paramount.

Retrieval-Augmented Generation (RAG) offers a promising remedy to these challenges. By augmenting generative models with external retrieval mechanisms, RAG grounds model outputs in factual evidence, thereby improving accuracy, contextual relevance, and trustworthiness[15-17]. While RAG has been successfully applied in open-domain question answering and certain clinical NLP applications, its potential in regulatory science - specifically for processing patient-generated device-related narratives - remains underexplored. Applying RAG to this setting is challenging due to the linguistic diversity and noise inherent in patient reports, but it holds the potential to transform the use of unstructured surveillance data.

Although zero-shot prompting and RAG are well-established methods in artificial intelligence (AI) research[18,19], their adoption in healthcare has primarily centered on structured or semi-structured data sources. For instance, retrieval-augmented strategies have been applied to improve interpretability and accuracy in zero-shot predictions on structured clinical tabular data[20], and to assess medical fitness using large-scale language models with retrieval grounding[21]. Similarly, another study has extracted highprecision relations from biomedical documents by reframing tasks as question-answering and using external biomedical knowledge to improve reasoning and reduce hallucinations[22]. However, these studies mostly operate in controlled settings, emphasizing structured datasets or well-defined benchmark tasks.

In contrast, patient-generated narratives in the FDA MAUDE system present a unique and underexplored challenge: they are highly unstructured, linguistically diverse, and inherently noisy. Existing NLP methods struggle with synonymy, negation, and implicit mentions, often leaving critical adverse event signals undetected. By leveraging RAG-enhanced LLMs for mesh implant surveillance, this study addresses this crucial gap. This approach introduces new challenges related to scalability, reliability, and interpretability, while directly advancing regulatory science by systematically extracting adverse event data from previously underutilized patient-reported sources. Importantly, our evaluation demonstrates not only accuracy improvements over conventional approaches but also statistically validated gains in adverse event detection, underscoring the clinical significance of this framework.

This study specifically analyzes patient-reported mesh implant narratives from the FDA MAUDE database (2000-2021) to transform unstructured reports into structured clinical insights. Our focus is on extracting key metadata (e.g., report date, hernia type, adverse event classification) and identifying both explicit and implicit adverse events. By integrating zero-shot prompting with a RAG framework, we aim to create a scalable, interpretable, and reliable pipeline for post-market surveillance of mesh implants.

This paper presents the following key contributions:

1. We develop and validate a RAG framework tailored for FDA MAUDE mesh implant narratives. This framework is capable of converting free-text patient reports into structured fields, which are essential for surveillance and regulatory analysis.

2. We apply zero-shot prompting to extract and classify adverse events from unstructured narratives, demonstrating improved reliability, reduced hallucination, and superior scalability compared with conventional NLP methods.

METHODS

Dataset and Benchmark Baseline

The dataset for this research was obtained from the FDA MAUDE database (https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm), comprising voluntary reports submitted by patients who underwent mesh implant surgery between 2000 and 2021[3,23]. After filtering for patient-generated narratives,

we selected a total of 2,422 mesh-related reports for analysis. Each report contained original structured fields, including Report Date, Report ID, and Adverse Event Flag (Yes/No), as well as an unstructured free-text narrative field (FoiText) that captures detailed patient descriptions of their post-surgical experiences. A representative example of these data fields is shown in Table 1.

The free-text patient narratives describe a wide array of subjective experiences and complications. While these are valuable for post-market surveillance, they present challenges for automated processing due to their linguistic complexity and variability. Detailed preprocessing steps, data extraction methods, and cleaning procedures were performed, as described extensively in our prior publication[24]. The preprocessing pipeline involved several key steps to ensure data quality and consistency. Raw FDA MAUDE reports were initially filtered to retain only patient-generated narratives, excluding submissions from manufacturers and healthcare providers. Text cleaning procedures included the removal of misspelled words, and the correction of common spelling variations in patient-reported symptoms. Reports with insufficient narrative content (fewer than 10 words) or missing critical metadata were excluded. Additionally, duplicate reports were identified and removed based on content similarity and temporal proximity.

In the previous study, we employed a combination of named entity recognition (NER), keyword extraction, and hierarchical stochastic block modeling (hSBM)-based topic modeling techniques to systematically extract and create two additional structured data columns from these same 2,422 reports: hernia type classification and detailed adverse event categorization[25]. The hernia types were categorized into five groups: inguinal, umbilical, incisional, SUI, and an unknown category for reports that did not explicitly specify the surgical site. Adverse events were grouped into 20 clinically relevant categories, including pain, infection, urinary issues, bowel problems, device-related complications, and others, as outlined in Table 1.

To provide a rigorous evaluation framework for the current study's RAG model, we utilized the structured outputs from our previous research as the Benchmark Baseline, which serves as the reference standard for assessing model performance. This Benchmark Baseline represents ground-truth labels that were independently generated using different methodological approaches (NER, keyword extraction, and topic modeling), along with expert evaluations, providing a robust foundation for assessing the accuracy, consistency, and clinical relevance of the AI-derived results. For the current evaluation, we employed stratified random sampling to select 200 reports from the original 2,422 reports, ensuring proportional representation across hernia types and adverse event categories. Importantly, these 200 evaluation reports were completely independent from the RAG model development process and remained unseen by the model during any phase of system design or prompt engineering. The remaining 2,222 reports formed the retrieval corpus accessible to the RAG framework during inference, enabling the model to provide relevant context for generating structured summaries and extracting adverse events.

The evaluation methodology employed a hold-out validation approach, where model performance was assessed by directly comparing the RAG-generated outputs with the corresponding Benchmark Baseline labels for the same 200 reports. No cross-validation was employed, as this study focuses on zero-shot evaluation against pre-established reference standards, rather than model training or fine-tuning. By comparing the structured outputs generated by the model - such as summaries containing report date, hernia classification, and extracted adverse events - we assessed the accuracy, consistency, and clinical relevance of the AI-derived results while ensuring complete independence between the label creation process and the RAG evaluation framework.

Table 1. Sample mesh implant report with structured fields

Field	Example value or description
Report ID	123456
Report date	YYYY-MM-DD
Flag adverse event	Y/N
Foi text (patient narrative)	I felt something bulging in my belly button, causing agonizing discomfort. The umbilical hernia surgery was supposed to help, but instead, it made things worse. My belly button swelled, and the pain in my abdomen intensified. To make matters worse, I experienced vomiting and skin rashes, adding to the misery.
Hernia types[18]	Incisional, umbilical, inguinal, SUI, unknown
Adverse events[18]	Pain, bowel problem, device problem, sexual problem, infection, urinary problem, stress, chronic inflammation, incontinence recurrence, nausea, prolapse recurrence, swelling, mesh erosion, bleeding, discomfort, sleeping issue, constipation, diarrhea, UTI, allergic reaction

ID: Identity of the report; SUI: stress urinary incontinence; UTI: urinary tract infection.

An important aspect of the benchmark comparison is its ability to systematically identify potential inaccuracies or hallucinations in the model-generated outputs. Specifically, we defined two types of errors: Type 1 (Misinterpretations) - where the model's output deviates from or misrepresents the original input, and Type 2 (Unsupported Outputs) - where the model generates information not found in the input narratives. These distinctions were crucial for thoroughly assessing the reliability and fidelity of the RAG-enhanced outputs.

To evaluate these aspects rigorously, we conducted zero-shot experiments using two model variants: a standard generative language model and a version enhanced by the RAG retrieval component. As illustrated in Figure 1, the prompt-based instruction approach guided the model in generating structured summaries by populating predefined output templates with information explicitly found in the narrative texts. The model was specifically instructed to leave any template fields blank if the input narratives lacked the relevant information, thereby minimizing speculative or unsupported generation and ensuring the outputs remained closely aligned with the patient-reported experiences.

RAG framework

To enhance the model's ability to generate accurate and contextually grounded outputs, we implemented a RAG framework tailored for analyzing unstructured patient narratives related to mesh implant surgeries. This approach combines information retrieval with generative modeling, enabling the language model to produce more reliable and evidence-backed summaries and extractions[26].

The workflow begins with the segmentation of lengthy FDA mesh reports into smaller, semantically meaningful text chunks. This step ensures that each text unit can be efficiently encoded and searched while preserving contextual integrity. In parallel, structured data - such as report metadata - is processed row by row to maintain alignment with the unstructured narrative.

Each text segment is then embedded using a sentence-transformer model from Hugging Face, which converts natural language into dense vector representations that capture semantic meaning and inter-word relationships[27]. These embeddings, along with associated metadata (e.g., Report ID, date), are stored in a vector database to enable high-speed, similarity-based retrieval. To identify the most relevant context for each query, we employ the Maximum Marginal Relevance (MMR) algorithm via the LangChain framework[28]. MMR was selected for its ability to balance relevance and diversity when selecting retrieved segments. For this study, we used a fixed parameter value of $k = 20$, consistent with common MMR practices[28], which balances retrieval comprehensiveness with computational efficiency. This setting allows

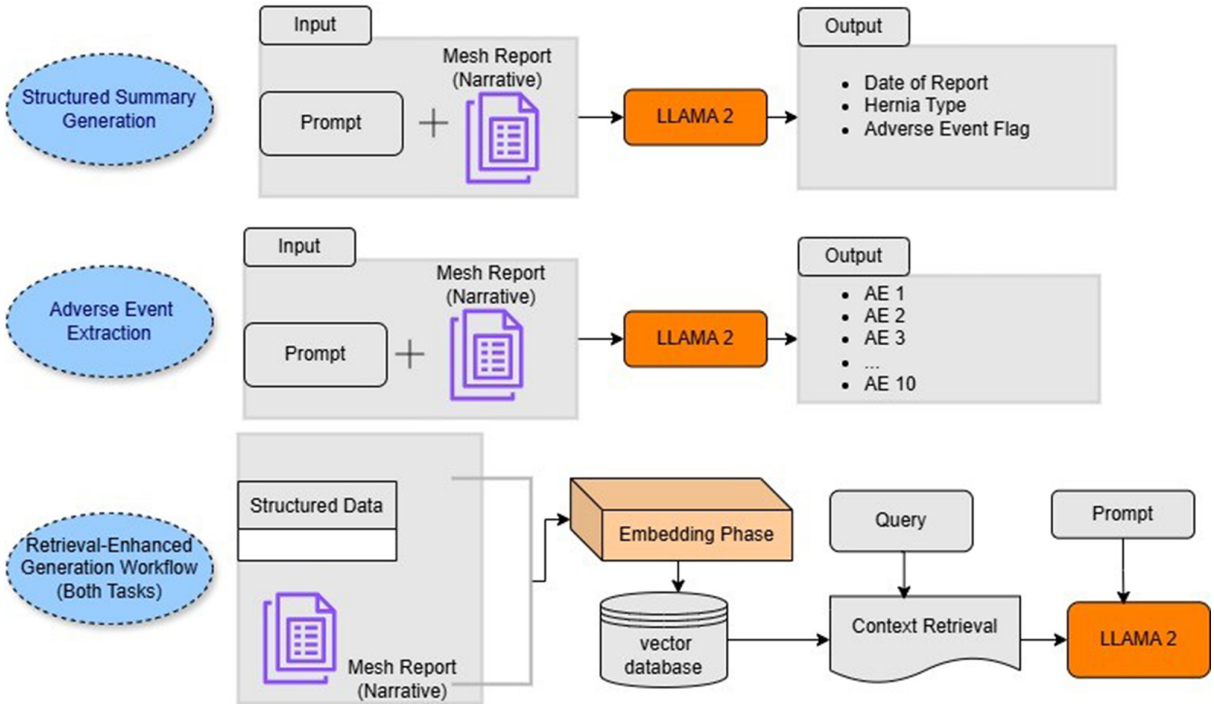

Figure 1. Overview of the model utilizing instruction tuning and RAG. RAG: Retrieval-Augmented Generation. (GitHub.com, https://github.com/InduBala-Y/RAG/blob/main/MESH_RAG.drawio.pdf)

the system to return a well-rounded set of informative and non-redundant text trunks. The retrieval process accepts parameters such as Report ID, Report Date, and a task-specific query (e.g., extracting hernia type or listing adverse events), and returns the most semantically aligned passages from the vector store. These retrieved segments are then concatenated with a task-specific prompt template and passed to the generative model, enabling it to produce structured summaries or adverse event lists grounded in the retrieved content. As illustrated in Figure 1, this process is integrated into both the structured summarization and adverse event extraction pipelines.

For the generative backbone, we utilized the LLaMA 2 model (13-billion parameter version) developed by Meta AI[29]. This open-source model was selected for its strong performance, accessibility, and ability to run locally, ensuring that sensitive patient data remain secure. The model weights were acquired through authorized access from Meta and Hugging Face repositories. To optimize computational efficiency, we implemented 4-bit quantization using the BitsandBytes library, substantially reducing the computational load while preserving generation quality.

The entire RAG pipeline - encompassing embedding, retrieval, and generation - was orchestrated using the LangChain framework, facilitating seamless integration between components. This RAG-enhanced approach substantially improved the fidelity and interpretability of outputs across both structured summarization and adverse event extraction tasks, as it provided the model with direct access to relevant patient-reported context. A detailed overview of the framework is provided in Figure 1.

RESULTS

Performance of structured summary generation

We evaluated the performance of a zero-shot structured summarization approach applied to FDA mesh implant reports. Using prompt-based instructions, the model was tasked with extracting key clinical variables from patient-written narratives, specifically focusing on Report Date, Hernia Type, and Adverse Event Flag. The model's outputs were compared against a Benchmark Baseline, which served as the reference standard for evaluating model performance in terms of both accuracy and clinical validity.

To quantify how well the model's output aligned with the Benchmark Baseline, we used the Jaccard Similarity Coefficient. This metric is particularly effective in scenarios involving multi-label classification, as it accounts for partial overlaps[30,31]. Given that individual reports may contain several adverse events - such as pain, vomiting, or mesh migration - the Jaccard score helps capture the degree of agreement between the model's extractions and those previously labeled. The Jaccard coefficient is calculated as the size of the intersection divided by the size of the union of the two sets.

The Jaccard similarity is defined as:

$$Jaccard\ (A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where A is the set of adverse events extracted by the model, and B is the corresponding set in the Benchmark Baseline. Beyond accuracy, we also examined the reliability of the model's outputs by categorizing its errors into two types[32]:

- Type 1 (Misinterpretation): The model generates content based on the input but distorts or misrepresents it.

- Type 2 (Unsupported output): The model generates content that is not supported by the input text at all.

To illustrate these error categories, consider the following examples from our evaluation dataset. For Type 1, a patient narrative stating 'I experienced discomfort after my hernia surgery in 2019, but the pain started getting worse in 2021' resulted in the model incorrectly extracting the report date as 2019 instead of 2021, demonstrating misinterpretation of temporal references. For Type 2, a narrative describing 'I felt something bulging and had severe discomfort after mesh surgery' led the model to generate adverse events including 'mesh migration' and 'infection', which were not mentioned in the original text, representing hallucinated content.

This dual-error framework allowed us to assess not only extraction accuracy but also the trustworthiness of the outputs - critical factors in clinical NLP tasks where incorrect interpretations can lead to misleading conclusions. A template of Structured Summary output generated from FDA reports using zero-shot prompting is provided in Table 2.

To evaluate model performance empirically, we randomly selected 200 patient reports from the FDA dataset and compared the structured outputs produced by the model to those in the Benchmark Baseline. Accuracy was measured at the field level, with a prediction counted as correct only if it exactly matched the corresponding label. Partial, inferred, or hallucinated responses were marked as incorrect unless they were clearly supported by the input text. To determine whether the observed differences between the RAG model and the Benchmark Baseline were statistically significant, we applied a Chi-square test for independence to

Table 2. Structured summary output generated from FDA reports

Structured summary template	Structured_summary template = Date of report: Type of hernia: Adverse event flag:
Prompt template	Prompt_template = [INST]<<SYS|||<*** You are an intelligent system designed to assist with structuring clinical reports related to hernia mesh implants. <</SYS|||<*** Using the given context, extract the following information: - Date of the report (if available) - Type of hernia (if explicitly mentioned): choose from "inguinal", "umbilical", "incisional", "SUI", or say "not mentioned" - Indicate whether the report clearly describes an adverse event ("Yes" or "No") Do not infer missing information. Leave any field blank if not directly stated in the context. context: [FDA_mesh report]
Input	[ID] I had been struggling with urinary incontinence for years that started in 2014. Despite initial hesitation, my doctor assured me the mesh was safe. After surgery, I experienced severe pain, difficulty during intercourse, and a feeling of something protruding. This has left me deeply uncomfortable.
Output	Date of report: Type of hernia: SUI Adverse event flag: Yes

FDA: Food and Drug Administration; ID: identity of the report; SUI: stress urinary incontinence.

each evaluated field[33].

All statistical analyses, including computation of accuracy, confidence intervals, Jaccard similarity, and Chi-square tests, were performed in Python 3.10 using the pandas, NumPy, and SciPy (stats) libraries. Visualization and tabulation were generated using Matplotlib and Seaborn to ensure analytical transparency and reproducibility.

For each report, we compared the number of correct and incorrect extractions produced by both models against the corresponding benchmark labels. The statistical analysis was carried out for key fields, including hernia type, adverse event flag, date of report, and adverse event extraction, with significance set at $P < 0.05$. This allowed us to assess whether improvements in accuracy were likely due to the RAG integration rather than random variation. The calculated *P*-values are presented in Table 3, alongside accuracy metrics and common error types.

The results in Table 3 demonstrate clear performance differences between models with and without RAG integration. Classification of the Adverse Event Flag achieved accuracies of 83% without RAG and 99% with RAG. The Report Date field showed accuracies of 67% without RAG and 81% with RAG. Hernia Type classification demonstrated accuracies of 60% without RAG and 82% with RAG. The Jaccard Overlap score for adverse event extraction showed values of 75% without RAG and 88.9% with RAG.

Extracting adverse events from patient reports

For this task, the model was prompted to extract up to 10 adverse events per patient report. A sample template report is presented in Table 4. During the extraction process, we observed that the model often captured either the full list of relevant events or a partial subset, depending on how clearly those events were described in the input narrative. To systematically evaluate performance, a total of 200 patient reports were analyzed. Half of these reports contained clearly described adverse events, while the remaining half included ambiguous phrasing or lacked explicit mentions of events.

Table 3. Comparison of structured information extraction performance with and without RAG

Field evaluated	Accuracy without RAG (%, 95% CI)	Accuracy with RAG (%, 95%CI)	Common error	*P*-value (0.05) (exact)
Report date	67 (60.1-73.5)	81 (75.1-86.2)	Misinterpretation (Type 1), Unsupported output (Type 2)	> 0.05 (0.062)
Flag adverse events	83 (79.4-87.8)	99 (96.4-99.9)	Misinterpretation	< 0.05 (0.001)
Hernia type	60 (56.1-66.7)	82 (76.1-87.2)	Misinterpretation, missing context	< 0.05 (0.032)
Adverse events	60 (58.4-65.4)	86 (80.7-90.5)	Partial event capture	< 0.05 (0.018)
Jaccard overlap (event match)	75 (69.7-80.7)	88.9 (84.1-92.8)	Misinterpretation (Type 1), Unsupported output (Type 2)	< 0.05 (0.025)

RAG: Retrieval-Augmented Generation.

Table 4. Adverse event extraction template from FDA reports

Adverse events template	adverse_events_template = AE 1: AE 2: AE 3: AE 4: ... AE 10:
Prompt template	prompt_template = [INST]<<SYS|||<*** You are an intelligent system designed to extract adverse event information from patient reports. <</SYS|||<*** Using the context provided, identify up to ten adverse events (each in four words or fewer) related to pain, discomfort, reactions, or device complications. If no adverse event is clearly mentioned, respond with "no event mentioned." Do not include inferred or speculative information. context: [FDA_mesh_report] adverse events template: [adverse_events_template] filled template: [/INST]
Input	[ID] had severe pain, not been able to eat, vomiting. I suspect that the mesh migration is adding to my issues. Disturbing sensation of something moving inside.
Output	AE 1: severe pain AE 2: vomiting AE 3: mesh migration AE 4: no event mentioned ... AE10: no event mentioned

FDA: Food and Drug Administration; AE: adverse event; ID: identity of the report.

In the 100 reports with clearly articulated adverse events, the model correctly identified the relevant events in 93 cases, resulting in an accuracy of 93%. Errors in the remaining 7 reports typically involved missing one or more relevant events due to subtle linguistic cues, paraphrased terms, or long and complex sentences. These were classified as partial captures rather than hallucinations. For the 100 reports with vague, implied, or non-standard event descriptions, the model adhered to prompt instructions correctly in 78 cases, identifying only events explicitly supported by the text. However, in 7 instances, the model generated adverse events not present in the original report (Type 2 hallucinations), and in 15 reports, it failed to identify events that were subtly embedded in the language, resulting in a 78% accuracy rate for this group.

Combining both subsets, the overall extraction accuracy was 86%, reflecting instruction compliance and contextual understanding. Comparison with the non-RAG model revealed consistently lower performance and a higher frequency of hallucinated responses, as detailed in Table 3.

Prevalence measurement of adverse events

To evaluate the model's ability to capture population-level trends in adverse event reporting, we compared the prevalence rates of each adverse event category extracted by the RAG model with those in the Benchmark Baseline. For this analysis, we focused exclusively on clinically meaningful reports - specifically, those that contained at least one adverse event label in the Benchmark Baseline. Reports without any adverse event annotations were excluded to ensure a more targeted and interpretable comparison.

As illustrated in Figure 2, the model identified Pain as the most frequently reported complication, with a higher prevalence than that observed in the Benchmark Baseline. This was followed by increased detection of other high-frequency adverse events such as Bowel Problems and Device Problems. The model also demonstrated greater sensitivity in identifying adverse events such as Urinary Problems, Swelling, and Mesh Erosion. In contrast, the model's output showed minimal overestimation for lower-prevalence complications such as Allergic Reaction and UTI, closely aligning with baseline estimates. Statistical comparison of prevalence rates revealed that the differences between the model outputs and the Benchmark Baseline were significant for several adverse event categories $P < 0.05$.

DISCUSSION

The results demonstrate that RAG integration significantly enhances the accuracy and reliability of structured information extraction from unstructured patient narratives. The substantial improvements observed across all evaluation metrics support the effectiveness of combining zero-shot prompting with retrieval augmentation for clinical text analysis.

Performance improvements and clinical significance

The most notable improvement was observed in the Adverse Event Flag classification, where accuracy increased from 83% to 99% with RAG integration. This near-perfect performance is particularly significant for clinical surveillance applications, where accurate identification of adverse events is critical for patient safety monitoring. The improvement suggests that RAG's contextual enrichment enables the model to better interpret vague or ambiguously stated symptoms that would otherwise be misclassified.

The Hernia Type classification also showed substantial improvement, from 60% to 82% accuracy, a 37% relative increase. This enhancement is clinically meaningful, as accurate hernia type identification is essential for understanding device performance across different surgical applications. The RAG model's ability to infer hernia types from contextual signals, even when not explicitly stated, demonstrates the value of retrieving relevant similar cases from the database. Report Date extraction improved from 67% to 81%, addressing a common challenge in clinical text processing, where temporal expressions can be ambiguous. The reduction in Type 1 misinterpretations indicates that RAG helps disambiguate between surgery dates, symptom onset dates, and actual report dates through contextual understanding.

Error analysis and model reliability

The dual-error framework (Type 1 and Type 2) provided valuable insights into model reliability. The reduction in both error types with RAG integration demonstrates improved factual grounding and reduced hallucination. Type 1 errors, primarily involving temporal misinterpretations, were significantly reduced through access to similar contextual examples. Type 2 errors, involving unsupported outputs, were minimized by grounding generation in retrieved relevant passages. The adverse event extraction task revealed notable patterns. The 93% accuracy for clearly articulated events versus 78% for ambiguous descriptions highlights the continued challenge of processing subjective patient language. Nonetheless, RAG maintained high precision while improving recall, offering a favorable trade-off for clinical applications.

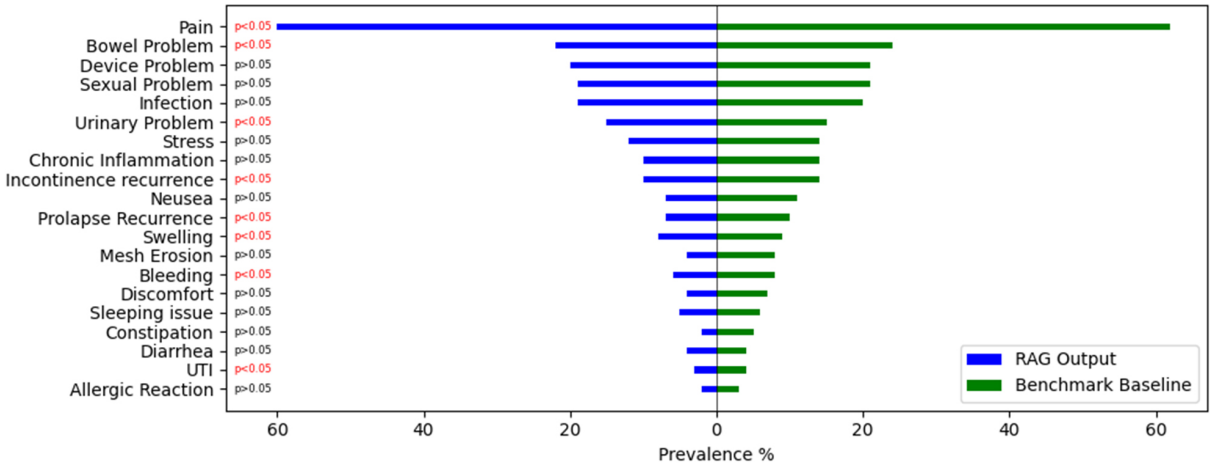

Figure 2. Prevalence comparison of adverse events identified by the RAG model and the Benchmark Baseline. RAG: Retrieval-Augmented Generation. (GitHub.com, https://github.com/InduBala-Y/RAG/blob/main/Stats%20test-RAG.png)

Population-level insights

Prevalence analysis revealed that the RAG model captured additional adverse event mentions that traditional keyword-based approaches might miss. Higher detection rates for common complications such as Pain, Bowel Problems, and Device Problems suggest that the model successfully identified subtly expressed or indirectly mentioned events. Importantly, precision was preserved for rare events, avoiding false-positive inflation that could mislead surveillance efforts. Statistically significant differences in prevalence rates between RAG outputs and the baseline indicate that the model may uncover previously under-detected adverse events in patient narratives. This finding has important implications for post-market surveillance, suggesting that current methods may underestimate the true prevalence of certain complications.

Methodological contributions

This study contributes to the growing body of work on RAG applications in healthcare by demonstrating its effectiveness in a specialized domain with high reliability requirements. The zero-shot approach eliminates the need for expensive domain-specific training data while achieving clinically relevant performance. The integration of MMR for diverse retrieval ensures that the model accesses varied contextual information, reducing the risk of bias from repetitive examples. Moreover, the framework's interpretability through retrievable source context addresses a critical need in clinical AI applications, where understanding the basis for AI decisions is essential for trust and validation.

In conclusion, this study presents the development of a RAG framework for extracting structured clinical information and adverse events from unstructured patient narratives related to mesh implant surgeries. By integrating zero-shot prompting with LLaMA 2 and retrieval-based context enhancement, the system achieved high accuracy in summarizing key report elements such as report date, hernia type, and adverse event flag. It also demonstrated strong performance in extracting multi-label adverse events with minimal hallucination and high fidelity to input narratives.

Evaluation against a Benchmark Baseline showed that the RAG-enhanced model consistently outperformed the non-RAG baseline across all key metrics. Statistical analysis using Chi-square tests further confirmed that improvements in accuracy were significant for most fields, especially for adverse event extraction and classification. Prevalence comparisons also demonstrated the model's ability to detect subtle or implicit

mentions of complications, offering a richer understanding of real-world patient outcomes.

However, this study has some limitations that should be acknowledged. First, the evaluation was confined to the FDA MAUDE dataset, without external validation across diverse clinical settings or international regulatory databases. Second, the k = 20 parameter for MMR was chosen based on common practices rather than systematic optimization, which may not be optimal for all clinical text types. Third, the study focused on English-language reports from a single regulatory system, limiting applicability to multilingual or international contexts. Finally, while RAG reduces hallucination, it cannot eliminate the inherent subjectivity of patient-reported narratives or guarantee complete accuracy in safety-critical applications, where human expert oversight remains essential.

Looking ahead, future work can explore fine-tuning the generative model on domain-specific corpora to further enhance contextual understanding. Incorporating external clinical ontologies or knowledge graphs could also support more precise identification and normalization of adverse event terms. Moreover, expanding the framework to handle multilingual reports or to integrate temporal progression of symptoms could provide deeper insights into long-term patient outcomes[34]. Lastly, combining this system with clinician feedback loops may support semi-automated adverse event validation pipelines for regulatory or pharmacovigilance purposes. Additionally, expanding validation beyond the FDA MAUDE dataset to include diverse clinical settings and direct engagement with healthcare stakeholders will be essential for establishing practical utility and supporting broader clinical adoption of this RAG-enhanced approach in healthcare decision-making systems.

DECLARATIONS

Authors' contributions

Conceptualized the study, developed the methodology, conducted the analysis, and wrote the initial draft: Bala I

Contributed to the development of the methodology and provided a critical review: Sharma E

Supervised the research, provided guidance throughout the study, and contributed to the review and editing of the manuscript: Mitchell L

All authors read and approved the final manuscript.

Availability of data and materials

The dataset used in this study is publicly available from the U.S. Food and Drug Administration's Manufacturer and User Facility Device Experience (MAUDE) database (https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm), which can be accessed online at the FDA website. The specific dataset and analysis codes are available from the corresponding author upon reasonable request.

Financial support and sponsorship

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interest

The authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Watson C. Hernia. In: Watson C, Davies J, Editors. Ellis and Calne's lecture notes in general surgery. Hoboken: John Wiley & Son; 2023. pp. 311-22. DOI
2. Le TN, Afshar Ali M, Gadzhanova S, et al. Hernia repair prevalence by age and gender among the Australian adult population from 2017 to 2021. *Critical Public Health.* 2024;34:1-11. DOI
3. Mishali M, Sheffer N, Mishali O, Negev M. Understanding variation among medical device reporting sources: a study of the MAUDE database. *Clin Ther.* 2025;47:76-81. DOI PubMed
4. Clavel M, Durán F, Eker S, et al. Maude manual (version 3.1). *SRI International*, 2020. Available from: https://gentoo.uls.co.za/distfiles/5d/Maude-3.1-manual.pdf [accessed 16 October 2025].
5. Kou Q, Wu M. Unlocking the potential of natural language processing in decoding medical device adverse events. In: Lane M, Sethumadhavan A, Editors. Collaborative intelligence: how humans and AI are transforming our world. Cambridge: MIT Press; 2024. pp. 197-211. DOI
6. I. Natural language processing in medical science and healthcare. *Medicon Med Sci.* 2023:4;1-2. DOI
7. Liao TJ, Crosby L, Cross K, Chen M, Elespuru R. Medical device report analyses from MAUDE: device and patient outcomes, adverse events, and sex-based differential effects. *Regul Toxicol Pharmacol.* 2024;149:105591. DOI PubMed
8. Bala I, Malhotra A. Fuzzy classification with comprehensive learning gravitational search algorithm in breast tumor detection. *IJRTE.* 2019;8:2688-94. DOI
9. Martin SC, Fitzgerald JJ. Tipu Zahed Aziz, MD (November 9, 1956-October 25, 2024). *Neuromodulation.* 2025;28:371-2. DOI
10. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. *arXiv* 2023; arXiv:2302.13971. Available from https://doi.org/10.48550/arXiv.2302.13971 [accessed 16 October 2025].
11. Bumgardner VK, Larsen MA, Anderson MB, Sayre GG, Fecho K, Pfaff ER. Local large language models for complex structured medical tasks. *arXiv* 2023; arXiv:2308.01727. Available from https://doi.org/10.48550/arXiv.2308.01727 [accessed 16 October 2025].
12. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ Digit Med.* 2024;7:16. DOI PubMed PMC
13. Zhang R, Han J, Zhou A, et al. LLaMA-Adapter: efficient fine-tuning of large language models with zero-initialized attention. *arXiv* 2024; arXiv:2303.16199. Available from https://doi.org/10.48550/arXiv.2303.16199 [accessed 16 October 2025].
14. Frayling E, Lever J, McDonald G. Zero-shot and few-shot generation strategies for artificial clinical records. *arXiv* 2024; arXiv:2403.08664. Available from https://doi.org/10.48550/arXiv.2403.08664 [accessed 16 October 2025].
15. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv* 2020; arXiv:2005.11401. Available from https://doi.org/10.48550/arXiv.2005.11401 [accessed 16 October 2025].
16. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. *arXiv* 2023; arXiv:2312.10997. Available from https://doi.org/10.48550/arXiv.2312.10997 [accessed 16 October 2025].
17. Salemi A, Zamani H. Evaluating retrieval quality in retrieval-augmented generation. In: SIGIR 2024: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024 Jul 14-18; Washington DC, USA. New York: Association for Computing Machinery; 2024. pp. 2395-400. DOI
18. Yu H, Guo P, Sano A. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In Machine learning for health (ML4H); 2023 Dec 10; New Orleans, USA. Cambridge: PMLR; 2023. pp. 650-63. Available from https://proceedings.mlr.press/v225/yu23b.html [accessed 16 October 2025].
19. Thompson WE, Vidmar DM, De Freitas JK, et al. Large language models with retrieval-augmented generation for zero-shot disease phenotyping. *arXiv* 2023; arXiv:2312.06457. Available from https://doi.org/10.48550/arXiv.2312.06457 [accessed 16 October 2025].
20. Mahbub S, Ellington C, Alinejad S, et al. From one to zero: RAG-IM adapts language models for interpretable zero-shot predictions on clinical tabular data. In: NeurIPS 2024 Third Table Representation Learning Workshop, 2024. Available from https://openreview.net/forum?id=3OYjWzqqC1 [accessed 16 October 2025].
21. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *NPJ Digit Med.* 2025;8:187. DOI PubMed PMC
22. Dong X, Zhao D, Meng J, Guo B, Lin H. SyRACT: zero-shot biomedical document-level relation extraction with synergistic RAG and CoT. *Bioinformatics.* 2025;41:btaf356. DOI PubMed PMC
23. Mishali M, Sheffer N, Mishali O, Negev M. Evaluation of reporting trends in the MAUDE Database: 1991 to 2022. *Digit Health.* 2025;11:20552076251314094. DOI PubMed PMC
24. Bala I, Kelly T, Stanford T, Gillam MH, Mitchell L. Machine learning-based analysis of adverse events in mesh implant surgery reports. *Soc Netw Anal Min.* 2024;14:1229. DOI

25. Boutin R, Bouveyron C, Latouche P. Embedded topics in the stochastic block model. *Stat Comput.* 2023;33:10265. DOI
26. S SK, G GJWK, E GMK, J MR, A RGS, E Y. A RAG-based medical assistant especially for infectious diseases. In: 2024 International Conference on Inventive Computation Technologies (ICICT); 2024 Apr 24-26; Lalitpur, Nepal. New York: IEEE; 2024. pp. 1128-33. DOI
27. Galli C, Donos N, Calciolari E. Performance of 4 pre-trained sentence transformer models in the semantic query of a systematic review dataset on peri-implantitis. *Information.* 2024;15:68. DOI
28. Wang X, Han Y, Tang M, Zhang F. Robust orbital game policy in multiple disturbed environments: an approach based on causality diversity maximal marginal relevance algorithm. In: Liu L, Niu Y, Fu W, Qu Y, Editors. Proceedings of 4th 2024 International Conference on Autonomous Unmanned Systems (4th ICAUS 2024); 2024 Sep 19-21; Shenyang, China. Singapore: Springer; 2025. pp. 355-69. DOI
29. Badshah S, Sajjad H. Quantifying the capabilities of LLMs across scale and precision. *arXiv* 2024; arXiv:2405.03146. Available from https://doi.org/10.48550/arXiv.2405.03146 [accessed 16 October 2025].
30. Verma V, Aggarwal RK. A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: empirical and theoretical perspective. *Soc Netw Anal Min.* 2020;10:660. DOI
31. Bala I, Kelly T, Lim R, Gillam MH, Mitchell L. An effective approach for multiclass classification of adverse events using machine learning. *JCCE.* 2024;3:226-39. DOI
32. Groves M, O'Rourke P, Alexander H. Clinical reasoning: the relative contribution of identification, interpretation and hypothesis errors to misdiagnosis. *Med Teach.* 2003;25:621-5. DOI PubMed
33. McHugh ML. The chi-square test of independence. *Biochem Med.* 2013;23:143-9. DOI PubMed PMC
34. Bala I, Mitchell L, Gillam MH. Analysis of voluntarily reported data post mesh implantation for detecting public emotion and identifying concern reports. *arXiv* 2025; arXiv:2509.04517. Available from https://doi.org/10.48550/arXiv.2509.04517 [accessed 16 October 2025].