Envisioning the Future of Health Informatics and Digital Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI250076

Leveraging LLMs to Understand Narratives in MAUDE Reports

Yuheng SHI[a] and Yang GONG[a,1]
[a] *University of Texas Health Science Center at Houston, Houston, Texas, USA*

Abstract. Interest in using the MAUDE database to investigate adverse events linked to medical devices has been growing. Yet, the narrative sections of these reports remain largely unexplored, leaving valuable insights unutilized and creating an incomplete understanding of these events. To bridge this gap, we employ large language models (LLMs) to analyze and interpret these narratives. Using OpenAI's GPT-4-turbo model, we focused on MAUDE reports involving endoscopic clips to identify uncoded surgical procedures and uncover additional insights. This approach showcases the potential of LLMs in processing narrative content, offering a more efficient and cost-effective alternative to previous methods and supporting the translation of MAUDE reports into actionable knowledge for clinical practice.

Keywords. MAUDE, large language models, narratives, endoscopic clips

1. Introduction

The FDA Manufacturer and User Facility Device Experience (MAUDE) is an open-access, valuable resource investigating adverse patient safety events associated with medical devices [1]. Recent studies have increasingly utilized MAUDE to explore adverse events across various medical fields, including endoscopy [2], ophthalmology [3], cardiology [4], and neurosurgery [5]. However, MAUDE medical device reports (MDRs) contain both categorical and textual fields, complicating the process of information retrieval and analysis. Quantitative analyses are commonly applied to structured data such as manufacturer names or generic names, while advanced unstructured data offer deeper context but require advanced processing techniques to extract meaningful insights. This dichotomy often leads to challenges in efficiently leveraging MAUDE data to comprehensively understand adverse events. To fully understand and learn from the events, it is essential to incorporate both categorical and textual fields in MDRs into the analysis.

Endoscopic events in the gastrointestinal (GI) tract have emerged as a key research area utilizing MAUDE. Endoscopic clips, specialized medical devices used for tissue localization, hemorrhage control, or securing equipment during endoscopic procedures, were chosen to illustrate our insights, building on previous studies and existing knowledge of related events [1]. However, structured data in MDRs often fails to capture the complexity and nuanced meaning embedded in narrative descriptions, particularly for GI endoscopic events, where understanding procedural context is crucial.

[1] Corresponding Author: Yang Gong, MD, PhD, FIAHSI. 7000 Fannin St, Ste 600. Houston, Texas, 77030, USA Email: Yang.Gong@uth.tmc.edu

Large Language Models (LLMs) have attracted significant attention for their remarkable abilities to process natural language in a human-like manner [6]. Leveraging these capabilities, LLMs are beginning to reshape biomedical fields by addressing challenges associated with electronic health records (EHRs), genomic sequences, medical imaging, and more [7]. To understand MDR narratives in the absence of standardized terminology, we leverage the contextual comprehension of LLMs, specifically GPT-4. Analyzing these MDR narratives allows us to transform MAUDE data into valuable knowledge, enhancing our understanding of incidents and improving clinical practices.

2. Methods

In our earlier projects on endoscopic clip reports, we analyzed MDRs from January 2012 to January 2021 [1], which serve as a foundation to explore MDR narratives. This study extends the analysis period to encompass the Q1 2021 as a pilot. To facilitate the collection of MDRs, we utilized OpenFDA, an open-source platform that provides access to raw MAUDE data [8]. The MDRs were systematically organized in a structured and searchable format, allowing us to conduct various keyword combinations. In this pilot, we utilized the medical specialty description of "Gastroenterology, Urology (GU)" alongside the Product Classification Code "PKL," which denotes a “hemostatic metal clip for the GI tract" [9]. Subsequently, we compared the MDR search results with those obtained through the previous search strategy using a different set of keywords. To ensure the robustness of our analysis, a threshold of 100 words in MDR narratives was established to exclude reports containing excessively brief narrative texts. The remaining reports were then converted into CSV format to facilitate further analysis.

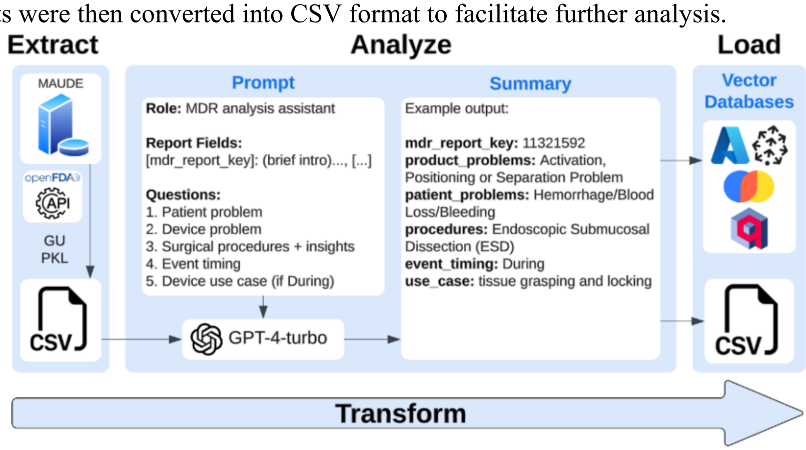

Figure 1. A workflow to extract-transform-load MDRs and analyze MDR narratives.

Prompt engineering techniques were applied to design the input query, assigning GPT the role of an assistant and providing a detailed overview of each field within the MDRs [10]. **Figure 1** depicts the workflow to analyze narratives in MDRs. We deployed LangChain, a software framework streamlining LLM applications, to implement automated analysis on the collected MDRs. GPT-4-turbo ("GPT") was selected as the model for this task, with the temperature parameter set to 0.0 to minimize the randomness of output. GPT was programmed to process one report at a time, focusing on five specific questions per report. The first two questions focused on the device issues and patient outcomes, leading GPT to elaborate on the "what" and "how" aspects while extracting relevant information from the text. The subsequent questions revolved around surgical procedures: Question 3 asked GPT to pinpoint specific surgical terminology and offer additional insights from the text; Question 4 requested GPT to classify the timing of adverse events as 'before', 'during', or 'after' the procedure; and Question 5 examined the device usage when the event occurred. GPT was instructed to present its responses in JSON format, with each response delivered as a string value. The output was stored in a single CSV file containing the report key and columns corresponding to responses for the five focused questions. The responses to Questions 3–5 were reviewed, and variated medical expressions were integrated to facilitate subsequent frequency analysis.

3. Results

A total of 95 endoscopic clip MDRs were extracted and added to our previously collected data, identifying 21 unique device problem categories and 12 patient problem categories. **Table 1** shows a comparison of search results from the prior search strategy and the current approach for endoscopic clip reports. The combination of medical specialty and product code resulted in a reduced number of device problems (N=159), while an identical number of patient adverse events was presented (N=12).

Table 1. A comparison of search results using different keyword combinations.

Search Fields	N_R	N_D (Average*)	N_P (Average*)
Manufacturer Name/Generic Name	102	172 (1.69)	12 (0.12)
Medical Specialty/Product Code	95	159 (1.67)	12 (0.13)

N_R: number of reports; N_D: number of device problems; N_P: number of patient adverse events.
* Average number of problems contributed per report: N / N_R

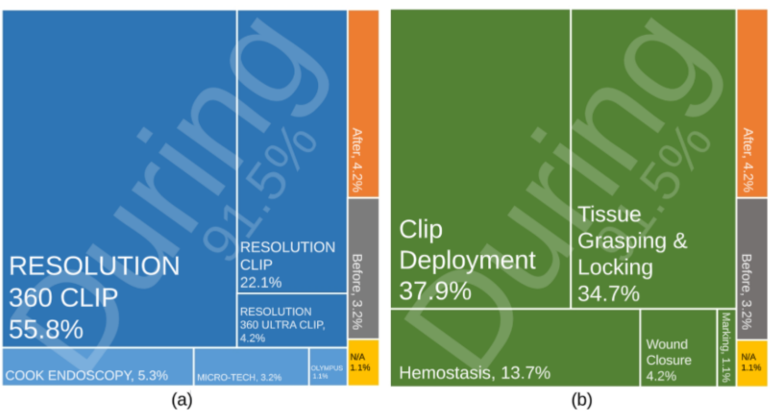

Figure 2. Treemaps contrasting endoscopic clip manufacturer distribution (a) in blue and use cases during procedures (b) in green.
Note: dark blue represents Boston Scientific Corp. products, and light blue indicates other manufacturers.

From the responses to the first two questions, the most common patient adverse events identified were "Hemorrhage", "Pain", and "Foreign Body in Patient", followed by "Perforation" and "Pancreatitis." For device-related issues, "Activation, Positioning, or Separation" was the most frequently reported problem, followed by "Mechanical Error (including break, jamming, material deformation, etc.)" and "Difficulty Open or Close."

GPT feedback for Question 3 identified nine specific surgical procedures, excluding miscellaneous and unspecified categories. "Colonoscopy" was the most frequently mentioned term in endoscopic clip MDRs (N=25), followed by "Esophagogastroduodenoscopy (EGD)" (N=24) and "Endoscopic Submucosal Dissection (ESD)" (N=9). For the event timing, GPT answers for Question 4 revealed that most adverse events were reported to occur during the procedure (N=87), with only three reports indicating the event occurred during preparation or beforehand. As a result, we undertook a detailed breakdown analysis of intra-procedural events (Question 5), illustrated through two tree maps in **Figure 2**. (a) shows most events are linked to clips manufactured by Boston Scientific Corp., with over half involving the Resolution 360 Clip. (b) depicts the typical use cases of clips when events occur, with "Clip Deployment" and "Tissue Grasping & Locking" being the most common, followed by "Hemostasis".

4. Discussion

Building on our previous work with ETL pipelines [1], this article addresses the lack of analysis of narrative data from MDRs. By utilizing the capabilities of GPT to comprehend and generate content, we are able to extract valuable, unclassified, procedure-specific information from contextual report data. This enables us to accurately identify surgical procedures and to reveal supplementary insights regarding events that previous MAUDE studies did not capture. This process is expected to seamlessly integrate into the proposed ETL pipeline, offering an efficient and accessible approach to transforming narratives into meaningful clinical knowledge.

The conventional MAUDE search methodologies predominantly concentrate on structured, searchable fields, thus underutilizing valuable information in MDR narratives. The surgical procedures and event timing identified in the narratives reveal the hidden gems of MAUDE. MDRs with product code "PKL" include procedures "Endoscopic Submucosal Dissection (ESD)" and "Endoscopic Mucosal Resection (EMR)," both widely recognized in endoscopy. Incorporating surgical terms with other searchable fields gives users flexible access to MDRs, capturing a broader spectrum of cases and enhancing insights into MAUDE.

LLMs have demonstrated considerable potential across a multitude of biomedical data types; however, their application to patient safety content signifies a novel utilization. Our results indicate that leveraging GPT in analyzing MDRs can be an efficient and accessible approach for extracting insights requiring less manual effort. In addition, integrating LLMs for narratives can amplify the clinical impact of MDR analyses, ultimately enhancing patient safety research.

The number of patient and device problems identified using the modified search strategy reveals a more significant variation in device problem categories compared to patient problem categories. This observation suggests that the defined categories in MAUDE contain redundant variations, complicating efficient labeling of reported issues and reducing the overall quality and clarity of MDRs. Additionally, unspecified or unknown categories undermine the clinical learning value of MDRs, making it challenging to derive clear insights for improving patient safety. We believe leveraging LLMs can help scrutinize problem categories by consolidating semantically overlapping terms and excluding unspecified cases, thereby streamlining the data-cleaning process and enhancing the dataset's consistency and accuracy.

The introduced LLM data analytical process aims to provide a reliable approach that frontline clinicians can quickly adopt. Although the demonstration was confined to the Q1 2021 dataset, extending the duration of collected MDRs is warranted, as this would facilitate a holistic view of the narratives. Furthermore, advanced database solutions such as vector stores could replace the basic CSV format, improving dataset read-write capabilities, preserving information integrity, and enabling better data management and flexible information retrieval. Additionally, the 100-word MDR threshold may inadvertently exclude concise, high-quality reports. As such, future studies should carefully revise the standards for filtering low-quality MDRs. Finally, the performance of GPT could be further optimized by employing techniques like fine-tuning or few-shot learning.

5. Conclusion

The current undertaking advances the efforts to analyze the MDR narratives by utilizing the progressively evolving LLMs. This process highlights the considerable significance of LLMs in extracting valuable information from MDR narratives, including surgical procedures and event timing, thereby augmenting the efficiency and learning value of MDRs to enhance patient safety studies based on the open-access patient safety datasets. Furthermore, it illustrates the potential for integrating LLMs into established workflows to facilitate the analysis of MDRs, transforming raw data into actionable insights and practical knowledge.

References

[1] Shi Y, Yu Y, Feng Y, Gong Y. A Data Pipeline for Enhancing Quality of MAUDE-Based Studies. Stud Health Technol Inform. 2024;316: 1214–1218. doi:10.3233/SHTI240629.

[2] Kumar VCS, Aloysius M, Aswath G. Adverse events associated with the gold probe and the injection gold probe devices used for endoscopic hemostasis: A MAUDE database analysis. World J Gastrointest Endosc. 2024;16:37–43. doi:10.4253/wjge.v16.i1.37.

[3] Hong S, Yu A, Wong A, Adamovich-Zeitlin R, Shah P, Minaya J, et al. Adverse Events of Femtosecond-Assisted Laser-Assisted In situ Keratomileusis: A Manufacturer and User Facility Device Experience Database Study. J Curr Ophthalmol. 2023;35:350–354. doi:10.4103/joco.joco_164_23.

[4] Litwinowicz R, Batko J, Rusinek J, Olejek W, Rams D, Kowalewski M, et al. LARIAT or AtriClip: Complications Profile and Comparison in Patients with Atrial Fibrillations Based on Manufacturer and User Facility Device Experience Database. Medicina (Lithuania). 2023 Nov 21;59(12):2055. doi:10.3390/medicina59122055.

[5] Porwal MH, Kumar D, Thalner S, Hedayat HS, Sinson GP. Analysis of reported adverse events of pipeline stents for intracranial aneurysms using the FDA MAUDE database. J Cerebrovasc Endovasc Neurosurg. 2023 Sep;25(3):275-287. doi: 10.7461/jcen.2023.E2022.10.010.

[6] Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. 2023 Nov 22;25(1):bbad493. doi: 10.1093/bib/bbad493.

[7] Nerella S, Bandyopadhyay S, Zhang J, Contreras M, Siegel S, Bumin A, et al. Transformers and large language models in healthcare: A review. Artif Intell Med. 2024 Aug;154:102900. doi: 10.1016/j.artmed.2024.102900.

[8] OpenFDA. OpenFDA [Internet]. 2024 [cited 2024 Nov 14]. Available from: https://open.fda.gov/.

[9] U.S. Food and Drug Administration. Product Classification [Internet]. 2024 [cited 2024 Nov 14]. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?id=2215.

[10] Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023 Oct 4;25:e50638. doi: 10.2196/50638.