# Bridging Gaps with Multimodal Data: A Comprehensive Dataset for Pharmacovigilance Analysis in Ovarian Cancer

**Anonymous ACL submission**

## Abstract

Ovarian cancer is a highly fatal type of gynecologic cancer, with over 70% of cases diagnosed at an advanced stage due to mild and nonspecific symptoms. This delayed diagnosis involves intensive treatments, such as surgery and chemotherapy. These treatments widely use platinum-based compounds and taxanes, which are highly effective but can cause serious adverse reactions. Identifying adverse drug reactions (ADRs) efficiently is essential in managing these side effects and ensuring that patients receive the most effective and safest medical care possible. In this work, we present *OvaCer*, a novel multi-labelled multimodal dataset thoroughly developed for ovarian cancer pharmacovigilance. This dataset includes 1500 records containing vital details such as drug name, duration of drug use, adverse effects, severity levels, post-effect actions, and reference images used during ovarian cancer treatment. In order to further enhance its adaptability for pharmacovigilance objectives, we have incorporated gold-standard summaries of patient experiences. Recognizing the potential of large language models (LLMs) in summarization, we conducted a comprehensive evaluation of several pre-trained models, including GPT-3.5, T5, BART, FlanT5, and clinical models like PMC LLaMA in medical summarization. Our results show that LLMs demonstrate varying degrees of effectiveness in clinical summarization tasks, with GPT-3.5 significantly outperforming other models.

## 1 Introduction

Ovarian cancer is ranked as the third most frequently diagnosed type of gynecologic cancer worldwide and appears to be a significant public health issue (Momenimovahed et al., 2019). It remains the leading cause of gynaecological cancer-related deaths in developed countries (Kurnit et al., 2021a). Despite advancements made in treatment methods, this disease continues to have a high mortality rate, with more than 70% of patients relapsing within the first five years after being diagnosed (Kuroki and Guntupalli, 2020; Stewart et al., 2019; Kurnit et al., 2021b). **Pharmacovigilance** is the scientific study and set of actions focused on finding, evaluating, understanding, and preventing any harmful effects or other issues related to drugs. The majority of ovarian cancer cases are detected at an advanced stage, necessitating aggressive treatment methods that are frequently toxic. Adverse drug reactions (ADRs) are common in oncology, with approximately 10-20% of cancer patients experiencing severe ADRs that require medical intervention. Chemotherapy drugs used to treat ovarian cancer, such as platinum-based compounds and taxanes, are known to have serious side effects. Effective pharmacovigilance can help to reduce ADRs, improve treatment adherence and outcomes, and lower hospitalization rates.

***Impact of research*** : *Pharmacovigilance studies have important implications in the field of ovarian cancer, as they address the widespread problem of under-reporting adverse drug reactions. Physicians often prioritize drug efficacy, sometimes overlooking ADRs as normal occurrences. Proactive pharmacovigilance enhances spontaneous reporting, which is crucial for gathering critical ADR information. These insights can prompt competent authorities to make informed decisions about each drug, such as discontinuing use, adjusting dosages, or taking other necessary steps that significantly improve treatment outcomes, benefiting society by raising the standard of ovarian cancer care.*

Furthermore, pharmacovigilance agencies utilize surveillance systems like FAERS (Li et al., 2014) to monitor drug safety post-market, but these systems face challenges such as under-reported and delayed data collection (Sarker et al., 2015). Manual data collection also hinders clinical evidence gathering for pharmacovigilance (Thompson et al., 2018). To address these issues, our research introduces *OvaCer* to streamline data availability for pharmacovigilance in ovarian cancer treatment. To sum up, our **key contributions** include:

- We introduce *OvaCer*, the first multi-labeled multimodal dataset for ovarian cancer, aimed at enhancing pharmacovigilance research and cancer care.
- We gather detailed annotations to provide specific and broad information about patients and conditions.
- We comprehensively evaluate pre-trained Large Language Models (LLMs) like GPT-3.5, T5, BART, FlanT5, and clinical models like PMC LLaMA to assess their effectiveness and limitations in medical summarization tasks.

## 2 Related Works

**Pharmacovigilance in Oncology**: In recent years, the detection and assessment of drug reactions associated with cancer treatments have drawn a lot of attention because of their potential impact on patient safety and treatment outcomes. While anticancer drugs have been thoroughly researched and proven to be highly effective in cancer treatment, they should be used with caution due to their high toxicity and narrow therapeutic window (Gandhi et al., 2005). Although these drugs effectively target and treat a variety of cancers, they also carry the risk of adverse drug reactions, which can range from mild and manageable to severe and require hospitalization (Shaikh and Nerurkar, 2022). A 2010 review of 95 articles identified that inaccurate reporting of adverse events could lead to more hospitalizations (Leendertse et al., 2010). Adverse Drug Reactions (ADRs) in oncology are common and often predictable, making them an essential part of the treatment process (Lau et al., 2004). However, it is common for oncology ADRs to go unreported because the adverse effects are often considered inevitable (Baldo and De Paoli, 2014). According to a few studies, follow-up calls can be effective in collecting information about adverse events (Monestime et al., 2021) and managing symptoms. However, there is limited evidence on the efficacy of follow-up calls for identifying adverse events that were not reported to a healthcare provider (Salmany et al., 2018; Spoelstra, 2017; Eldeib et al., 2019).

Nevertheless, in recent years there has been significant progress in the accurate reporting of adverse drug reactions in oncology. Furthermore, the deployment of digital pharmacovigilance systems has the potential to improve cancer patients' quality of life by facilitating the timely reporting of adverse reactions (Salathé, 2016; Khozin et al., 2017). Scientific societies are also making significant progress toward developing guidelines, tools, and platforms for reporting ADRs in clinical trials and oncology research (Absolom et al., 2017; Levit et al., 2018).

**Clinical Datasets**: The current datasets, such as the PSB 2016 social media shared task dataset (Sarker et al., 2016), the Medline ADE corpus (Gurulingappa et al., 2012), the CADEC dataset (Karimi et al., 2015), and the BioDEX dataset (D'Oosterlinck et al., 2023), consist of adverse drug events (ADEs) across a wide range of clinical fields. This indicates a significant gap in datasets designed specifically for monitoring ADEs in cancer treatment. To address this limitation, we introduce our dataset specific to **OVA**rian can**CER**, *OvaCer*, which consists of ADEs associated with anticancer drugs used in ovarian cancer treatment.

## 3 Corpus Development

The literature review highlights that previous research, while substantial, has significant gaps in addressing oncology-related pharmacovigilance, particularly for ovarian cancer. To address this gap, we have developed a novel dataset *OvaCer* developed to support a variety of tasks related to ovarian cancer pharmacovigilance. We have provided different statistics for the *OvaCer* dataset in Table 1. The steps we took to prepare this corpus are listed below.

| **Measures** | **Size** |
|---|---|
| *No. of Samples* | 1500 |
| *Number of True labels (Adversity)* | 1141 |
| *Number of unique Drugs reported* | 109 |
| *Number of distinct effects reported* | 532 |
| *Number of images* | 400 |

Table 1: Statistics of *OvaCer* Dataset

### 3.1 Data Collection

A recent qualitative analysis of online discussion forums was conducted to investigate the perspectives of ovarian cancer patients regarding ADEs caused by anticancer medications. A thorough online search was carried out to identify relevant internet forums. We identified the Cancer Survival Network (CSN)[1] public healthcare blog for its open access and active patient involvement in side effects and treatment.


Figure 1: An instance of adverse event caused by drugs used in ovarian cancer treatment

### 3.2 Data Annotation

To ensure comprehensive and ethical annotation, we enlisted two medical students and one Ph.D. student, each meeting specific criteria: a minimum age of 25 years, fluency in English, and a willingness to handle sensitive content. Participants were compensated for their involvement, and the annotation process was completed within four months. To verify the quality of the annotated data, we established rigorous standards that each sample had to meet:

- For each post mentioning multiple drugs and numerous effects (positive and negative), extract only those drug names linked to adverse drug events (negative effects).
- Each data instance's adversity of the drug event is assessed using specific terms indicating adversity, such as "bad," "worse," "unbearable," "irrecoverable," "permanent," or similar expressions conveying similar sentiments.
- Each data instance's severity of the drug event is assessed based on explicit mentions of congenital anomalies, life-threatening situations, disabilities, or hospitalizations (initial or prolonged). If these criteria are not explicitly stated, the severity is categorized as not applicable to that specific data point.
- Reference images illustrating physical effects experienced by patients under similar drug treatments are added to each relevant data instance as depicted in Figure 3. Instances not related to drug side effects are removed.
- Every data point includes a URL link. For each data instance, access the content at that URL to gain insight and context about the data.

To maintain consistency among annotators, final labels were assigned via majority voting. Annotators were instructed to remain objective without bias related to demographics or other factors. To enhance our dataset for pharmacovigilance applications, we created detailed summaries for each post, including relevant details such as medicinal needs, disease, drug names, disorders, symptoms, and age. We thoroughly evaluated the summaries produced by our method using several reading scores, like abstractness, concreteness, Flesch-Kincaid grade, Dale-Chall readability score, and Coleman-Liau index demonstrated in Table 2. A detailed explanation for these parameters is provided in the APPENDIX A.2. This evaluation ensures that the summaries accurately represent the original posts and are understandable to readers of varying linguistic abilities.

| ***Metrics*↓** | ***OvaCer*** |
|---|---|
| *Concreteness* | 0.772 |
| *Flesch Kincaid Grade* | 12.366 |
| *Dale Chall Score* | 11.476 |
| *Coleman Liau Index* | 14.043 |
| *Number of samples* | 1500 |

Table 2: Readability scores used to assess the Gold standard summaries for *OvaCer* dataset.

## 4 Models

In our work, we assessed the performance of several standard summarization models, including T5 (Vaswani et al., 2017), BART (Lewis et al., 2019), GPT 3.5 (Brown et al., 2020), FlanT5 (Chung et al., 2022), and some clinical models, namely PMC Llama (Wu et al., 2023), on the *OvaCer* dataset. These models were chosen due to their remarkable performance in various summarization datasets in recent years, as demonstrated by previous studies (Laskar et al., 2022; Ravaut et al., 2022).

**T5**: An adaptable transformer-based model (Vaswani et al., 2017) utilizes a single text-to-text transfer learning framework to handle multiple tasks, including translation, summarization, and question-answering.

**BART**: A transformer-based sequence-to-sequence model pre trained for document

[1]https://csn.cancer.org/

denoising (Lewis et al., 2019).

**FlanT5 small**: (Chung et al., 2022) Flan-T5 Small is an improved version of the T5 model (Vaswani et al., 2017), fine-tuned for various text-to-text NLP tasks such as summarization and translation with reduced computational resources.

**PMC Llama**: (Wu et al., 2023) PMC-LLaMA is the first open-source language model specifically designed for medical applications. It incorporates data-centric knowledge and is fine-tuned with medical-specific instructions.

## 5 Experimental Results and Analysis

To evaluate the model-generated summaries against gold reference summaries, we used ROUGE scores (Lin, 2004) and BERTScore (BS) (Zhang et al., 2020). Rouge-1 measures unigram overlap, indicating the summary's relevance; Rouge-2 assesses bigram overlap, reflecting coherence; Rouge-L evaluates the longest common subsequence, indicating structural accuracy; and BERTScore uses BERT embeddings to assess semantic similarity. Detailed explanations of these evaluation metrics can be found in the APPENDIX A.1 section. These metrics collectively provide a comprehensive assessment of the model‘s performance in capturing relevant information, maintaining coherence, and ensuring semantic accuracy. The results of our evaluation, as demonstrated in Table 3, indicate that GPT-3.5 outperforms other models on all metrics, demonstrating its efficiency and capability in medical summarization. It excels with a high R-1 score, effectively capturing essential single words, and a high R-2 score, demonstrating proficiency in understanding bigram relationships. The R-L score reflects consistent coherence in sentence structure when compared to reference summaries, whereas the BS score reflects strong semantic similarity, indicating a firm grasp of context and meaning. The T5 model performs fairly well but lags significantly behind GPT-3.5. The R1 score indicates a moderate ability to capture unigrams, while the lower R2 score indicates difficulty in accurately capturing bigrams. However, the BS score for the T5 model suggests sufficient semantic understanding with some potential for improvement. In comparison to T5, BART exhibits lower performance across all metrics. It struggles with both unigram and bigram capture, as indicated by lower R-1 and R-2 scores, and shows weaker coherence in summaries based on the R-L score. Additionally, BART's BS score suggests less semantic alignment with reference summaries. Similarly, Flan T5 also faces challenges with unigram and bigram capture, reflected in its low R-1 and R-2 scores. While it maintains reasonable semantic alignment, indicated by its comparable BS score to T5, Flan T5 encounters difficulties in maintaining coherent sentence structures, as indicated by its R-L score. PMC LLaMA shows poor results across all metrics. This indicates that these models are not suitable for summarizing clinical posts. The extremely low R-1, R-2, and R-L scores indicate significant difficulties in capturing n-gram models and producing coherent, relevant, and accurate summaries. This evaluation highlights the efficacy of GPT-3.5 for medical summarization tasks and emphasizes the necessity for strong models to handle the complexity of clinical text summarization effectively.

| **Models ↓** | **R-1** | **R-2** | **R-L** | **BS** |
|---|---|---|---|---|
| *GPT-3.5* | **0.461** | **0.186** | **0.309** | **0.896** |
| *T5* | 0.265 | 0.097 | 0.196 | 0.859 |
| *BART* | 0.238 | 0.065 | 0.156 | 0.832 |
| *Flan T5* | 0.178 | 0.060 | 0.133 | 0.848 |
| *PMC LLaMA* | 0.134 | 0.011 | 0.090 | 0.828 |

Table 3: Quantitative evaluation using Rouge-1, Rouge-2, Rouge -L and BERT Score

## 6 Conclusion

Our research addresses the challenge of limited resources in the field of pharmacovigilance for ovarian cancer by introducing a multi-label, multimodal dataset, the *OvaCer*. This contribution includes a collection of 1500 records, each accompanied by summaries and relevant images. By continuously monitoring and analyzing ADR data, healthcare providers can make informed decisions about drug safety, dosage adjustments, and alternative treatments, resulting in more efficient and effective ovarian cancer treatment. Furthermore, inspired by advancements in large language models (LLMs), we have conducted a comprehensive evaluation to assess their summarization capabilities using zero-shot prompting techniques within the context of ovarian cancer pharmacovigilance, concluding that LLMs exhibit varying degrees of effectiveness in the clinical summarization task, with GPT-3.5 outperforming other models significantly.

## 7 Limitations

The limitations of our research primarily relate to the size of the sample and the size of the visual data included. Our dataset has a smaller sample size compared to other clinical datasets. Furthermore, the images in our dataset are limited to adverse drug events (ADEs) that appear on external body parts, such as skin rashes or swelling. This dataset does not include images depicting internal conditions such as neck pain, fever, or nausea.

## 8 Ethical Consideration

In healthcare summarization, ethical considerations such as safety, privacy, and bias are critical. During the curation of *OvaCer*, we strictly adhered to established legal, ethical and regulatory standards. Additionally, the dataset does not reveal user identities, thereby preserving privacy and confidentiality. The annotation guidelines were approved by two medical researchers from the oncology department and a medical practitioner from the pharmacology department. Furthermore, after the dataset curation was completed, it was verified and approved by these experts. To ensure compliance and ethical integrity, we also obtained formal approval from our institute's healthcare committee and ethical review board (ERB) before utilizing the dataset for research purposes.

**Intended Use** We make our dataset publicly available to encourage further research into ovarian cancer pharmacovigilance. The dataset is released exclusively for research purposes, and we do not grant licenses for commercial use.

## References


Kate Absolom, Patricia Holch, Lorraine Warrington, Faye Samy, Claire Hulme, Jenny Hewison, Carolyn Morris, Leon Bamforth, Mark Conner, Julia Brown, et al. 2017. E lectronic patient self-r eporting of a dverse-events: P atient i nformation and a d vice (erapid): a randomised controlled trial in systemic cancer treatment. *BMC cancer*, 17:1–16.

Paolo Baldo and Paolo De Paoli. 2014. Pharmacovigilance in oncology: evaluation of current practice and future perspectives. *Journal of Evaluation in Clinical Practice*, 20(5):559–569.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Karel D'Oosterlinck, François Remy, Johannes Deleu, Thomas Demeester, Chris Develder, Klim Zaporojets, Aneiss Ghodsi, Simon Ellershaw, Jack Collins, and Christopher Potts. 2023. Biodex: Large-scale biomedical adverse drug event extraction for real-world pharmacovigilance. *arXiv preprint arXiv:2305.13395*.

Hend K Eldeib, Maggie M Abbassi, Marwa M Hussein, Salem E Salem, and Nirmeen A Sabry. 2019. The effect of telephone-based follow-up on adherence, efficacy, and toxicity of oral capecitabine-based chemotherapy. *Telemedicine and e-Health*, 25(6):462–470.

Tejal K Gandhi, Sylvia B Bartel, Lawrence N Shulman, Deborah Verrier, Elisabeth Burdick, Angela Cleary, Jeffrey M Rothschild, Lucian L Leape, and David W Bates. 2005. Medication safety in the ambulatory chemotherapy setting. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 104(11):2477–2483.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Sean Khozin, Gideon M Blumenthal, and Richard Pazdur. 2017. Real-world data for clinical evidence generation in oncology. *JNCI: Journal of the National Cancer Institute*, 109(11):djx187.

Katherine C Kurnit, Gini F Fleming, and Ernst Lengyel. 2021a. Updates and new options in advanced epithelial ovarian cancer treatment. *Obstetrics & Gynecology*, 137(1):108–121.

Katherine C Kurnit, Gini F Fleming, and Ernst Lengyel. 2021b. Updates and new options in advanced epithelial ovarian cancer treatment. *Obstetrics & Gynecology*, 137(1):108–121.

Lindsay Kuroki and Saketh R Guntupalli. 2020. Treatment of epithelial ovarian cancer. *Bmj*, 371.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.

Phyllis M Lau, Kay Stewart, and Michael Dooley. 2004. The ten most common adverse drug reactions (adrs) in oncology patients: do they matter to you? *Supportive care in cancer*, 12:626–633.

Anne J Leendertse, Djurre Visser, Antoine CG Egberts, and Patricia MLA van den Bemt. 2010. The relationship between study characteristics and the prevalence of medication-related hospitalizations: a literature review and novel analysis. *Drug Safety*, 33:233–244.

Laura A Levit, Raymond P Perez, David C Smith, Richard L Schilsky, Daniel F Hayes, and Julie M Vose. 2018. Streamlining adverse events reporting in oncology: an american society of clinical oncology research statement. *Journal of Clinical Oncology*, 36(6):617–623.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Hui Li, Xiao-Jing Guo, Xiao-Fei Ye, Hong Jiang, Wen-Min Du, Jin-Fang Xu, Xin-Ji Zhang, and Jia He. 2014. Adverse drug reactions of spontaneous reports in shanghai pediatric population. *PLoS One*, 9(2):e89829.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zohre Momenimovahed, Azita Tiznobaik, Safoura Taheri, and Hamid Salehiniya. 2019. Ovarian cancer in the world: epidemiology and risk factors. *International journal of women's health*, pages 287–299.

Shanada Monestime, Ray Page, Nicole Shaw, Randy Martin, William Jordan, Jessica Rangel, and Subhash Aryal. 2021. Factors associated with adherence to follow-up calls in cancer patients receiving care at a community oncology practice. *Journal of Oncology Pharmacy Practice*, 27(5):1094–1101.

Mathieu Ravaut, Shafiq Joty, and Nancy F Chen. 2022. Summareranker: A multi-task mixture-of-experts reranking framework for abstractive summarization. *arXiv preprint arXiv:2203.06569*.

Marcel Salathé. 2016. Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health. *The Journal of infectious diseases*, 214(suppl_4):S399–S403.

Sewar S Salmany, Lujeen Ratrout, Abdallah Amireh, Randa Agha, Noor Nassar, Nour Mahmoud, Dalia Rimawi, and Lama Nazer. 2018. The impact of pharmacist telephone calls after discharge on satisfaction of oncology patients: A randomized controlled study. *Journal of Oncology Pharmacy Practice*, 24(5):359–364.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212.

Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2016. Social media mining shared task workshop. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 581–592. World Scientific.

Sana P Shaikh and Rajan Nerurkar. 2022. Adverse drug reaction profile of anticancer agents in a tertiary care hospital: An observational study. *Current Drug Safety*, 17(2):136–142.

Sandra L Spoelstra. 2017. Oral anticancer agents: an intervention to promote medication adherence and symptom management. *Number 2/April 2017*, 21(2):157–160.

Christine Stewart, Christine Ralyea, and Suzy Lockwood. 2019. Ovarian cancer: an integrated review. In *Seminars in oncology nursing*, volume 35, pages 151–156. Elsevier.

Paul Thompson, Sophia Daikou, Kenju Ueno, Riza Batista-Navarro, Jun'ichi Tsujii, and Sophia Ananiadou. 2018. Annotation and detection of drug effects in text for pharmacovigilance. *Journal of cheminformatics*, 10(1):1–33.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

# A Example Appendix

## A.1 Quantitative Scores

Below, we explain the quantitative measures used to compare the summarization with gold reference summaries.

- ROUGE-1 score: This score is used to evaluate the quality of text summarization or machine-generated text compared to a reference or gold standard summary considering unigrams.

- ROUGE-2 score: This score measures the overlap of bigrams (pairs of consecutive words) between the generated summary and the reference summary. This metric captures some level of fluency and coherence, as it considers pairs of words rather than individual words.

- ROUGE-L score: This score considers the longest common sequence of words in both the generated and gold standard summaries.

- BERT((Bidirectional Encoder Representations from Transformers) ) score: This score computes a similarity score based on contextual embeddings from the BERT model, capturing semantic similarity between the generated and reference text.

### A.2 Readability Scores

The readability scores used to assess the written summaries are explained below:

- Concreteness: The summary's utilization of specific details and language to express the original poem's ideas and imagery.

- Flesch-Kincaid Grade: Evaluating the Flesch-Kincaid Grade ensures that the summary is written at a suitable level of difficulty, making it accessible to a diverse audience.

- Dale-Chall Readability Score: This metric helps determine whether the summary is written clearly and straightforwardly, allowing for easy comprehension.

- Coleman-Liau Index: This metric provides insight into the summary's overall readability and syntactic complexity, allowing us to identify areas for improvement in clarity and readability.

### A.3 Dataset Samples


Figure 2: An instance of adverse event caused by drugs used in ovarian cancer treatment


Figure 3: An instance of adverse event caused by drugs used in ovarian cancer treatment