# Survival Analysis for Cancers of the Brain, CNS and Bone using Retrieval Augmented Generation on the SEER Database

**Jyothi Vaidyanathan[1], Shourya Gupta[2], Justin Lee[1], Srikanth Prabhu[2], Saptarshi Sengupta[1] ***

[1]San Jose State University, San Jose, CA, United States
[2]Manipal Institute of Technology, Manipal, Karnataka, India
{jyothi.vaidyanathan, justin.lee03, saptarshi.sengupta}@sjsu.edu, 13.shourya@gmail.com, srikanth.prabhu@manipal.edu

## Abstract

Mortality estimation remains a key issue in cancers affecting the Brain, Central Nervous System (CNS), and Bone, among others. The recent integration of LLM-based reasoning into tools that aid cancer prognosis has been particularly encouraging. This prompts us to examine further their stated efficacy and devise workarounds to reduce hallucinations using retrieval augmented generation. We study the clinical, pathological and demographic logs of patients recorded in the National Institutes of Health (NIH) Surveillance, Epidemiology, and End Results (SEER) database and develop an integrated methodology that is user-friendly and responds to n-shot queries with or without context. We first build a set of custom SEER embeddings using DistilBERT, which we use to test tree-based models in answering 'yes/no' type 5-year survivability questions given patient profiles. We extend the limited binary response capability of the prior models by using TabLLM, HyDE-RAG, and Step-Back RAG on the BCNS cancer data and extend them to Bone Cancer data from SEER using GraphRAG, as the attributes are similar. The conversation-friendly models are able to take different context lengths and types into account and provide reasoning about their responses. We successfully show that the extensive patient records in the SEER database can be utilized to develop a powerful conversational agent that is not only able to classify mortality outcomes but also reason about the response by leveraging latent inter-relationships among the unique clinical variables.

## Introduction

Despite advancements in detection and treatment, cancer's complexity challenges early detection, precise prognosis, and effective treatment. Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has the potential to optimize current oncological treatments and can revolutionize the healthcare industry.

In healthcare, input data encompasses diverse modalities, including text, images, videos, and multimodal data. Traditional machine learning approaches do not integrate the full spectrum of structured and unstructured medical data effectively, thereby failing to capture the intricate biological and clinical complexities of cancer. These findings underscore the critical necessity for advancing ML methodologies, leveraging state-of-the-art deep learning and multimodal data fusion techniques.

Large Language Models (LLMs) are an intriguing development in DL, capable of analyzing vast amounts of structured and unstructured data, such as Electronic Health Records (EHRs), genomic data, medical literature and data from Cyber-Physical Healthcare Systems (CPHS). They can provide task-specific results, enabling development of personalized treatment plans for patients, prognosis, and facilitate accurate and informed natural language interactions. Using data from Surveillance, Epidemiology, and End Results (SEER) (National Cancer Institute 1975-2021), we use Generative AI to predict the survivability of cancer patients suffering from brain and CNS cancers. The user interacts with a front-end chat user interface (UI) that ingests prompts and additional context to extract insights from the data. Building on the results obtained, we extend our work to effectively apply Generative AI for bone cancer prognosis.

Furthermore, LLMs facilitate the translation of intricate analytical outputs into comprehensible language. This capability enables healthcare professionals to effectively interpret findings without necessitating advanced technical expertise, thereby enhancing clinical decision-making. As part of this research, we develop LLM models to answer mortality prediction queries in SEER data.

## Prior Work

ML is increasingly used to develop predictive models for prognosis and treatment in healthcare. In (Yu et al. 2023), researchers built several models using the SEER data to forecast 5-year survival in non-metastatic cervical cancer patients, highlighting the ability to interpret complex medical data and improve clinical decision-making.

In (Nobin, Rahman, and Alam 2022), the SEER dataset is used by 10 traditional machine learning classifiers to predict the survivability of patients with tonsil cancer, where random forest classifier showed the highest results with 93.88% accuracy. Prediction of Lung metastases (LM) and three-month prognostic factors in hepatocellular carcinoma (HCC) patients is done by researchers in (Alkhawaldeh et al. 2023), where Random Forest, Artificial neural network and Easy Ensemble classifiers are utilized. The Easy Ensemble and Random Forest models demonstrated the best perfor-

*Corresponding author. Email: saptarshi.sengupta@sjsu.edu

mance in predicting outcomes. Both studies, however, encountered the challenge of imbalanced data and addressed this issue in different ways.

LLMs transform tabular data tasks through usage of NLP and traditional machine learning. Hegselmann et al. (2023) show that by serializing data into natural language, LLMs outperform deep learning methods in zero- and few-shot classification tasks due to their pre-existing knowledge base. The paper 'TABLET: Learning From Instructions For Tabular Data' (Slack and Singh 2023) shows that LLMs improve zero-shot performance by 44% with Flan-T5 and 13% with ChatGPT when given natural language instructions.

Many LLMs have been pre-trained in medical datasets specifically for medical applications, such as Me-LLaMA 13B and Me-LLaMA 70B (Xie et al. 2024). By continuous fine-tuning, they outperform other open-source medical LLMs like GPT-4(OpenAI et al. 2024). Google's Med-PaLM 2(Singhal et al. 2023a) is notable, achieving 86.5% accuracy on medical exam queries and delivering expert-level answers. (Singhal et al. 2023b). Another notable LLM model, OncoGPT(Jia et al. 2024), is fine-tuned for oncology-related predictions, which is able to leverage a dataset of more than 180,000 oncology-related notes.

The evaluation metrics for LLMs differ slightly from those of traditional machine learning models. In (Guinet et al. 2024), researchers have created an automated way of RAG(Lewis et al. 2021) evaluation on specific tasks using synthetic multiple choice questions, using item response theory (IRT). LLMs are difficult to evaluate due to the variety of capabilities and inadequate benchmarks as discussed by (Zheng et al. 2023).

## Our Approach

We use Large Language Models (LLMs) to explore the survivability prediction problem, subsequently selecting the one that provides the most accurate and contextually relevant responses to the user. To implement the prognosis, the following tools are used - Compute Unified Device Architecture (CUDA) enabled GPUs (T4 and A-100), Google Colab/Jupyter Notebook, Python, PyTorch, LangChain.

Data Acquisition and Pre-processing: The SEER dataset is a comprehensive and continuously updated collection of cancer-focused information from various registries throughout the United States. It includes various details about cancer patients, such as demographic, diagnostic, treatment, and outcome data. To enhance predictive power, we conducted a comprehensive review of existing literature, performed Exploratory Data Analysis (EDA), and applied categorical feature encoding for effective feature selection.

**The features used for the Survivability Task are:**

'Site recode ICD-O-3/WHO 2008', 'Patient ID', 'Age recode with single ages and 85+', 'CS version input original (2004-2015)', 'RX Summ–Surg Prim Site (1998+)', 'Year of diagnosis','ICD-O-3 Hist/behav', 'CS extension (2004-2015)', 'First malignant primary indicator', 'Grade (thru 2017)', 'CS version input current (2004-2015)', 'Primary Site', 'Laterality', '5 year survivability', 'Sex', 'Race/ethnicity', 'Median household income inflation adj to 2019'

**(1) TabLLM:** TabLLM(Hegselmann et al. 2023) is a model designed for few-shot classification of tabular data. It is used in conjunction with the SEER dataset in order to predict cancer survivability outcomes.

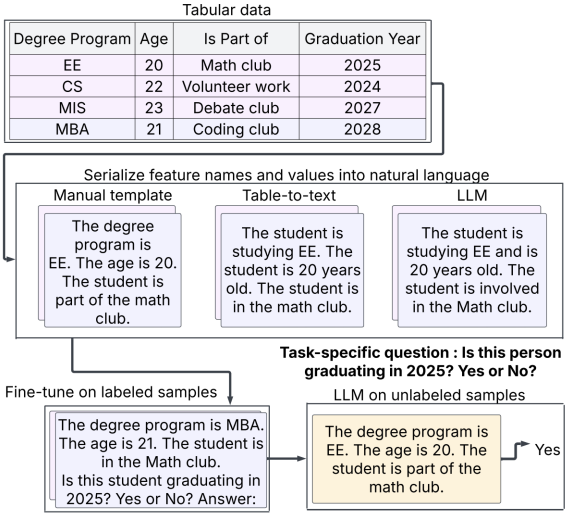


Figure 1: Overview of TabLLM (Hegselmann et al. 2023)

**Data Serialization:** Tabular data is serialized using templates converting rows to sentences. Consider a table with columns 'Degree Program', 'Age', 'Is Part of', and 'Graduation year'. A template converts a row like 20, Math club, 2025 into a sentence: 'A person aged 20 in the Electrical Engineering Program graduates in 2025.' (Figure 1).

**Prompting and Fine-Tuning:** Tabular data for prompting has been successfully serialized. The T0-11b model is employed to approach few-shot classification. It uses 4 labeled examples to fine-tune the model for the specific task: "Will this patient survive the next 5 years? Yes or no?".

**Design, Modeling and Module Specifications:** TabLLM's architecture is designed to enhance few-shot learning on tabular data through several key components. Tabular data are serialized for efficient data access and processing. Templates standardize query methods, enhance interpretability while scripts automate template creation, data preparation, and input compliance.

**(2) Use of Embeddings:** This methodology utilizes the capabilities of LLMs to transform structured data into numerical representations appropriate for traditional machine learning models. This approach uses embeddings processed from DistilBERT (Sanh et al. 2020) to train and supervise the various classification models.

**(3) Retrieval Augmented Generation (RAG):** It consists of two main components: a retriever and a generator. A retriever searches a large knowledge base for relevant information based on the user's prompt, while a generator uses the retrieved information to generate a contextually relevant response. This combination allows the LLM to incorporate external information in real-time. It is a superior approach to TabLLM, as it can retrieve updated data without retraining, unlike TabLLM, and can handle large unstructured datasets.

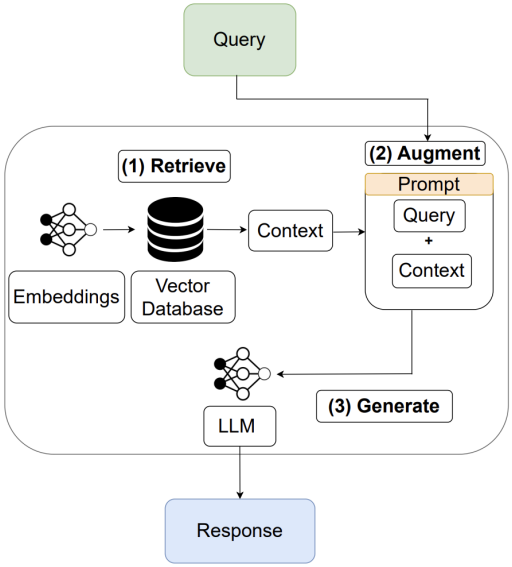


Figure 2: A General RAG Framework

For brain and CNS cancers, serialized representation of the database was fed as external context to the RAG models.

- **Data Preprocessing and Loading:** The PDF is loaded and processed using PyPDF2 to extract text which is concatenated into a single string. Patient information is segmented using regular expressions, and each segment corresponds to a distinct patient record.
- **Embedding Generation:** The *FastEmbedEmbeddings* model from LangChain is used to encode each segmented data as numerical vectors. The embeddings are stored in the Chroma(Chroma AI 2024) Vector Store, enabling efficient similarity searches based on user queries.
- **Retrieval Mechanism:** The Chroma Vector Store enables the retriever to identify similar patient records. For each query, the top 20 relevant records are retrieved, and the query is contextualized using a prompt template.
- **Generative Model Integration:** After retrieving records and reformulating queries, the Llama-3(Dubey et al. 2024) model is used to generate coherent answers. Prompt templates are useful for procedurally guiding generation and ensure relevant, structured responses.
- **Question Processing and Response Generation:** User queries are structured to be compatible with Llama-3 processing, allowing for the retrieval of relevant patient record segments. These embeddings are added to the RAG chain, which generates contextually aligned answers. *RunnablePassthrough* in LangChain integrates retrieval and generation components to streamline the process and deliver the final result.

**(3.1) RAG Variations and Optimizations:** Two variations of RAG were used: Step-Back, HyDE.

**RAG using Step-Back Prompting (Zheng et al. 2024)**: This technique iteratively refines user queries to enhance response accuracy and contextual alignment. It involves submitting a query, retrieving relevant data, generating a preliminary response, and refining the query through a feedback loop (Figure 4). Users can rephrase queries to extract additional insights to enhance the quality of the queries and the final output. LLama-3 is employed for response generation.

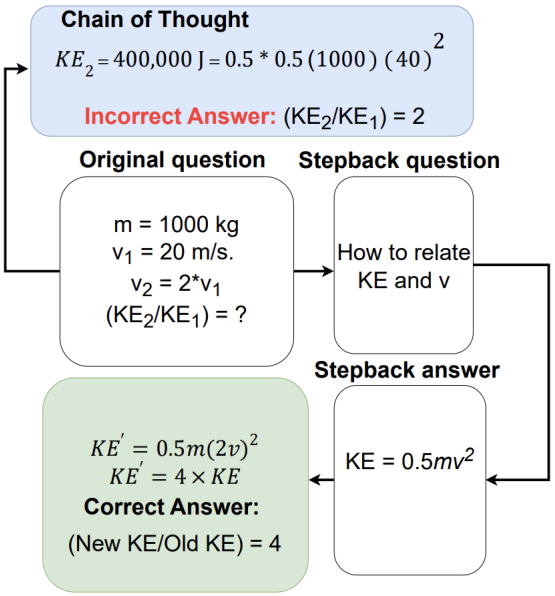


Figure 3: A Step-back prompting example showing an instance of a high-school physics concept (Zheng et al. 2024).

**HyDE-RAG:** Hypothetical Document Embeddings (Gao et al. 2022) method is used in RAG to improve the quality of document retrieval and response generation (Figure 4). By combining document retrieval and hypothetical reasoning, HyDE is an advanced and context-aware retrieval system compared to RAG. From a user's initial query, the LLM (e.g. GPT-3) generates a hypothetical response. The generated response replaces the original query as the input for the document retriever. The contextual information is transformed into a numerical vector, facilitating a similarity search within the vector store followed by retrieval.

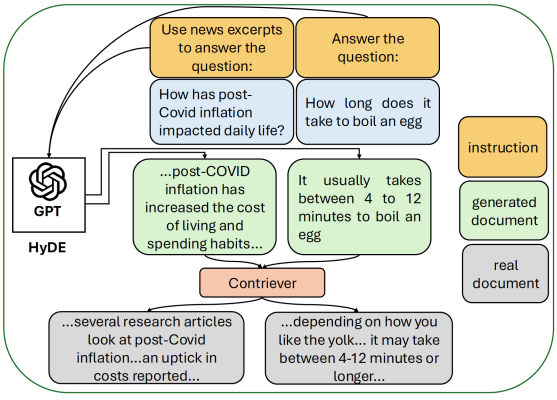


Figure 4: The HyDE model (Gao et al. 2022)

Building on these implementation techniques, we have been working on RAG approaches for Bone cancer prognosis using SEER data, starting with an approach called GraphRAG(Microsoft 2024).

**GraphRAG:** Unlike traditional RAG, which retrieves information based solely on vector similarity, GraphRAG utilizes LLM-generated knowledge graphs to establish connections between distinct data points through shared attributes. This structured approach enables a deeper contextual understanding. Particularly relevant to our research, it facilitates survivability prediction by enabling non-experts to leverage its capabilities without technical expertise.

## Experimental Results

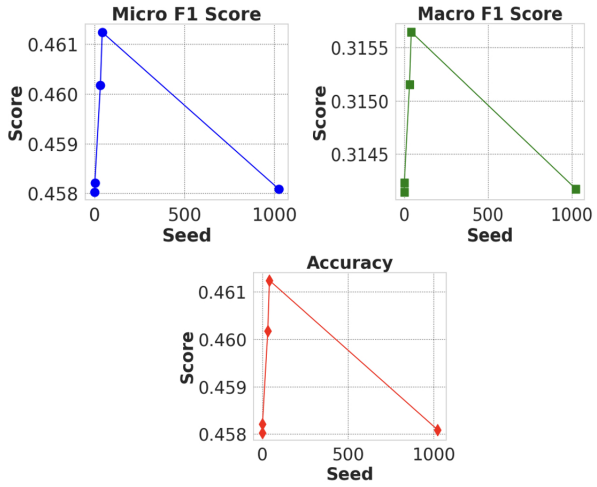


Figure 5: Results of various metrics for TabLLM (4 shots)

**TabLLM:** Figure 5 presents performance metrics, including Micro F1 Score, Macro F1 Score, and Accuracy, across various seeds and a shot count of 4 for the 5-year survivability task using the SEER dataset for TabLLM. Using 4 shots for TabLLM is motivated by the scarcity of labeled data in medical datasets. It serves as a baseline for understanding the model's data requirements, while also ensuring computational efficiency and enabling rapid experimentation. The model performed consistently on different seeds (seed 42 produced the most promising results), however, the baseline achieved with TabLLM was not optimal. It was observed that increasing the number of shots did not lead to a significant improvement in performance.

**DistilBERT Embeddings and Tree-based Models:** The DistilBERT language model was used to construct embeddings and evaluated using performance metrics, including Accuracy, Precision, Recall, F1 Score, and ROC AUC.

**Significance of the Results Obtained for the Survivability Task:** In predicting the 5-year survival rate for brain cancer patients, boosting methods such as Gradient Boosting and XGBClassifier outperform other techniques due to their ensemble nature, which enhances accuracy and reduces overfitting by iteratively correcting the mistakes of earlier trees. These models achieve higher ROC AUC values than others through effective handling of different data distributions, as shown in Figure 6. RandomForest performs well by averaging multiple decision trees, allowing it to balance variance and bias.

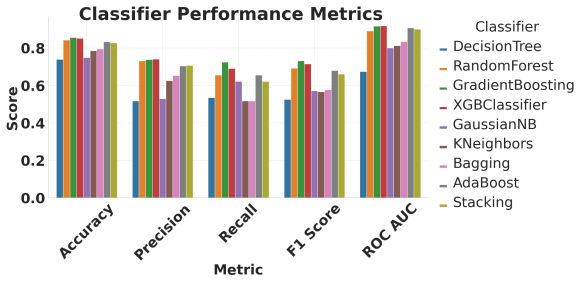


Figure 6: Survivability Task Comparative Analysis

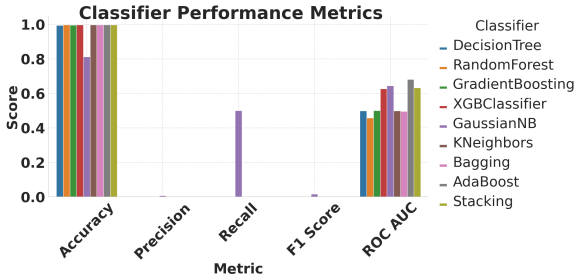


Figure 7: Metastasis Task Comparative Analysis

**Significance of Results for Bone Metastasis Task:** The results for Bone Metastasis highlight significant challenges due to the unbalanced dataset, as shown in Figure 7. The dataset has 3361 rows as 'No Metastasis' and only 6 rows as 'Yes Metastasis'. As a result, most classifiers demonstrate high accuracy but very low precision and recall. Many classifiers achieve ROC AUC scores close to 0.5, which suggests their performance is equivalent to that of random guessing. However, models like AdaBoost and XGBClassifier show better class discrimination with their ROC AUC values exceeding 0.6. Despite low overall performance, GaussianNB demonstrates the ability to identify the minority class.

**Retrieval Augmented Generation:** Two benchmarks were evaluated with Simple RAG to establish baseline performance on Llama-3. When tested using the context from the PubMedQA dataset, which provides 'yes' or 'no' answers based on biomedical literature, the model achieved 64.29% accuracy when answering questions about cancer. For domain-specific long-form cancer-related questions, without external context, the model achieved responses with an average cosine similarity of over 60%. This is promising, but there is also a need for refinement when handling detailed queries.

**Comparative Analysis of the RAG Variants:** Step-Back Prompting RAG achieves the highest accuracy at 0.70, excelling in overall correct classifications (Figure 8). Vanilla RAG has superior recall at 0.81 and the highest F1 score at 0.58. However, all methods struggle with low precision.

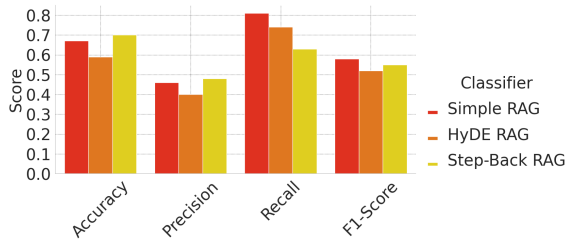


Figure 8: Comparative Analysis of RAG Variants

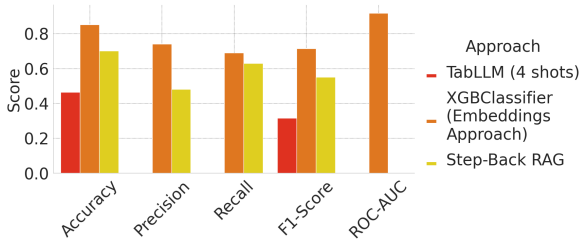


Figure 9: Comparative Analysis of Different Techniques

**Key Takeaways:** The results in Figure 9 indicate that XGBClassifier (Embeddings Approach) outperforms other models with high accuracy (0.8505), precision (0.7407), recall (0.6897), F1 score(0.7143), and ROC AUC(0.9177). The bar chart highlights XGBClassifier's superior performance. However, step-back prompting has broader applications due to its ability to handle diverse query encodings in prompts and interact directly with the user. SimpleRAG also provides the best precision-recall balance.

**Preliminary results on Bone Cancer using GraphRAG and Pinecone:** We implemented GraphRAG on the complete data for bone cancer logs in the SEER database. It provides a robust solution with detailed responses, especially when cost is not a limiting factor. We experiment with both global and local search options provided by GraphRAG using the prompt *'What are the common primary sites for a white male?'*. Figures 10-11 demonstrate that the local search method deliver more precise results for specific, narrowly-defined queries, making them suitable for use cases requiring tailored answers. We also implemented a chat application with a vector database Pinecone (Pinecone 2025). In Figure 12, significant cases can be attributed to patients in the mentioned age group.

## Conclusion and Future Work

For survival analysis in the Brain and CNS cancer patients, the ensemble XGBClassifier using custom SEER embeddings outperforms other models. However, step-back prompting allows for user interaction and enhances query encoding. For bone cancer, preliminary work indicates GraphRAG excels in delivering detailed, context-rich responses, enabling reasoning and precise information retrieval. These experiments demonstrate the feasibility of posing the cancer survival prediction problem as an LLM-driven natural-language understanding and response generation task, although rigorous benchmarking needs to be done to verify the usefulness, alignment and reliability of the said models. Generative models allow for detailed responses when working with large datasets and also enable integration with external sources to retrieve relevant answers if they are not available within the data. The next step is to evaluate responses and fine-tune. A comprehensive evalua-

**Prompt:**
What are the common primary sites for a white male?

**Response:**
In the context of bone cancer, particularly Osteogenic Sarcoma, the common primary sites for a white male include specific locations where the cancer manifests. The primary sites for Osteogenic Sarcoma in a white male often involve the long bones of the upper limb, scapula, and associated joints (C40.0), the long bones of the lower limb and associated joints (C40.2), the bones of the skull and face and associated joints (C41.0), the mandible (C41.1), and potentially other areas like the rib, sternum, clavicle, and associated joints (C41.3). These sites represent the areas where Osteogenic Sarcoma commonly affects white males, emphasizing the importance of understanding the specific body parts involved for accurate diagnosis and treatment planning. Moreover, the relationships within the Osteogenic Sarcoma community highlight that Osteogenic Sarcoma specifically affects white individuals, including white males, underscoring the demographic impact of this type of bone cancer. Understanding the demographic distribution of Osteogenic Sarcoma, especially in white males, is crucial for developing effective treatment and prevention strategies tailored to the specific needs of this demographic group. Overall, the primary sites commonly affected by Osteogenic Sarcoma in white males involve the long bones of the upper and lower limbs, the skull and face, the mandible, and areas like the rib, sternum, and clavicle. This information underscores the significance of recognizing these primary sites to enhance the management and outcomes of Osteogenic Sarcoma in white male individuals.

Figure 10: Local Search Prompt/Response with GraphRAG

**Prompt:**
What are the common primary sites for a white male?

**Response:**

### Common Primary Sites for Sarcomas in White Males

Analyst reports have highlighted that white males are commonly associated with various types of sarcomas, such as Osteogenic Sarcoma, Notochordal Sarcoma, Chondrogenic sarcoma, Fibrosarcoma, and Malignant Fibrous Histiocytoma. These sarcomas primarily affect specific anatomical structures within the body. The primary sites affected in white include the long bones of the lower limb, associated joints, bones of the skull and face, vertebral column, pelvic bones, sacrum, coccyx, ribs, sternum, and clavicle.

### Implications for Healthcare Interventions

Understanding the prevalence and impact of these sarcomas on white males is crucial for healthcare interventions, treatment planning, and improving healthcare outcomes within the community. The intersectionality of race and sex in sarcoma associations, particularly in reference to white males, underscores the intricate nature of health disparities faced by individuals belonging to multiple demographic groups. By recognizing the specific types of sarcomas linked to white males, such as Osteogenic Sarcoma and Notochordal sarcoma, healthcare providers may tailor interventions to address the unique health risks and disparities prevalent in this demographic group.

Overall, the primary sites commonly affected by Osteogenic Sarcoma in white males involve the long bones of the upper and lower limbs, the skull and face, the mandible, and areas like the rib, sternum, and clavicle. This information underscores the significance of recognizing these primary sites to enhance the management and outcomes of Osteogenic Sarcoma in white male individuals.

Figure 11: Global Search Prompt/Response with GraphRAG

| **Prompt:** |
| --- |
| I have a patient's data who is 63 years old, I do not know the gender of the patient though. As per records, which bone cancer(s) should I potentially get them tested for and why? |
| **Response:** |
| The patient, being 63 years old, could potentially be at risk for osteogenic sarcoma, other high grade sarcomas, or chondrogenic sarcoma based on the records provided. Osteogenic sarcoma and other high grade sarcomas were found in individuals around this age range in the data. Additionally, chondrogenic sarcoma was also present in patients within a similar age group. Therefore, it would be advisable to consider testing for these types of bone cancers in the patient. |

Figure 12: Prompt/Response using Pinecone

tion may be done using the 'LLM as a Judge' (Zheng et al. 2023) and multiple LLM models can be employed to ensure robustness in the assessments. Future work will aim at expanding the use of LLM-based CPHS to incorporate additional health indicators outside of what is available in the SEER data for cancers of the Brain, CNS and Bone.

## Acknowledgements

The research reported in this publication was supported by the Division of Research and Innovation at San José State University under Award Number 23-UGA-08-044. The content is solely the responsibility of the author(s) and does not necessarily represent the official views of San José State University.

## References

Alkhawaldeh, I. M.; Altarawneh, G.; Al-Jafari, M.; Abdelgalil, M. S.; Tarawneh, A. S.; and Hassanat, A. 2023. A Machine Learning Approach for Predicting Lung Metastases and Three-Month Prognostic Factors in Hepatocellular Carcinoma Patients Using SEER Data. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, 1–5.

Chroma AI. 2024. Chroma Documentation: Getting Started. Available: https://docs.trychroma.com/docs/overview/getting-started. Accessed: Feb. 07, 2025.

Dubey; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gao, L.; Ma, X.; Lin, J.; and Callan, J. 2022. Precise Zero-Shot Dense Retrieval without Relevance Labels. arXiv:2212.10496.

Guinet, G.; Omidvar-Tehrani, B.; Deoras, A.; and Callot, L. 2024. Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation. arXiv:2405.13622.

Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, 5549–5581. PMLR.

Jia, F.; Liu, X.; Deng, L.; Gu, J.; Pu, C.; Bai, T.; Huang, M.; Lu, Y.; and Liu, K. 2024. OncoGPT: A Medical Conversational Model Tailored with Oncology Domain Expertise on a Large Language Model Meta-AI (LLaMA). arXiv:2402.16810.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Microsoft. 2024. GraphRAG Documentation. Available: https://microsoft.github.io/graphrag/. Accessed: Feb. 7, 2025.

National Cancer Institute. 1975-2021. Cancer statistics. https://seer.cancer.gov/data/. Accessed: 2025-02-07.

Nobin, R. H.; Rahman, M.; and Alam, M. J. 2022. Survivability Prediction for Patients with Tonsil Cancer Utilizing Machine Learning Algorithms. In *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 210–215.

OpenAI; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pinecone. 2025. Pinecone. https://www.pinecone.io. Accessed: 2025-02-07.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; and et al.., D. N. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Slack, D.; and Singh, S. 2023. Tablet: Learning from instructions for tabular data. *arXiv preprint arXiv:2304.13188*.

Xie, Q.; Chen, Q.; Chen, A.; Peng, C.; Hu, Y.; Lin, F.; Peng, X.; Huang, J.; Zhang, J.; and et al.., V. K. 2024. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*.

Yu, W.; Lu, Y.; Shou, H.; Xu, H.; Shi, L.; Geng, X.; and Song, T. 2023. A 5-year survival status prognosis of non-metastatic cervical cancer patients through machine learning algorithms. *Cancer Medicine*.

Zheng, H. S.; Mishra, S.; Chen, X.; Cheng, H.-T.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2024. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. arXiv:2310.06117.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.