Endourological Society Student Scholarship Summary

Project Title: Development and Initial Proof of Concept of a Retrieval-Augmented Generation Large Language Model Decision Aid for the Surgical Treatment of Benign Prostatic Hyperplasia (RAPTOR-BPH-DA)

Student: Keiran Pace

Supervisor: Dr. Dean Elterman

Over the summer months, significant progress was made on the development and initial proof of concept of RAPTOR-BPH-DA, a next-generation, artificial intelligence–powered decision aid designed to enhance shared decision-making for the surgical treatment of benign prostatic hyperplasia (BPH). This work addressed a growing challenge in urologic care: the increasing number of surgical options for lower urinary tract symptoms secondary to BPH, each with differing risk profiles, functional outcomes, and patient considerations. Existing decision aids are often limited in scope, static in nature, and difficult to update, while general-purpose large language models (LLMs) such as ChatGPT are prone to inaccuracies and outdated information.

To overcome these challenges, key clinical guidelines from the Canadian Urological Association (CUA), American Urological Association (AUA), and European Association of Urology (EAU) were curated, cleaned, and formatted into a structured, hierarchical knowledge base. A novel recursive abstractive processing strategy, RAPTOR, was then applied to cluster and summarize semantically related guideline data, optimizing it for retrieval-augmented generation (RAG). This approach allowed the tool to dynamically retrieve and integrate relevant, evidence-based information at runtime, ensuring accurate and up-to-date responses to complex clinical questions.

The summer work focused on building this end-to-end framework, which included advanced techniques such as document chunking, semantic indexing, multi-query transformation, reciprocal rank fusion, and guardrail prompting to prioritize patient safety and readability. Once the tool architecture was established, it was benchmarked against standard ChatGPT performance using 88 domain-specific BPH questions. Evaluation metrics included F1 score, precision, and recall for accuracy, as well as blinded quality ratings by independent reviewers using a 5-point Likert scale and statistical testing with Wilcoxon signed-rank methods.

The completed work demonstrated that domain-specific RAG frameworks significantly outperform standard LLMs in delivering accurate, patient-appropriate, and guideline-based recommendations. This proof of concept provides a foundation for the next phases of the project, which will focus on usability testing with patients and clinicians, workflow integration, and ultimately, prospective studies of its impact on decision-making and clinical outcomes.