DOMAIN-SPECIFIC ARTIFICIAL INTELLIGENCE: VASC.AI'S IMPACT ON VASCULAR SURGERY AND EDUCATION

Tiam Feridooni[1,2,3], Arshia P. Javidan[1,2,3], Asha Behdinan[1,2,3], Daniyal N. Mahmood[1,2,3], Polycronis P. Akouris[4], Hirad Feridooni[5], Andrew Dueck[1], Mark Wheatcroft[2], David Szalay[2]

[1]Division of Vascular Surgery, Sunnybrook Health Sciences Center, University of Toronto, Toronto, ON, Canada

[2]Division of Vascular Surgery, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada.

[3]University Health Network, Peter Munk Cardiac Centre, Division of Vascular Surgery, University of Toronto, Toronto, Ontario, Canada.

[4]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[5]Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada

INTRODUCTION

The evolution of large language models (LLMs) has spurred significant interest in their clinical and educational applications. General-purpose models—such as ChatGPT (versions 3.5, 4, and 4o), Gemini, and Copilot—are trained on broad datasets, but they can suffer from "hallucinations" and inaccuracies when addressing specialized topics. In vascular surgery, where accurate diagnosis and tailored management are essential, a domain-specific solution is imperative. VASC.AI addresses these limitations by integrating Retrieval-Augmented Generation (RAG) with a curated database containing over 250,000 vascular surgery–related abstracts, clinical guidelines, and landmark trial data. Our research progressed in two phases. First, we conducted the VESAP study to evaluate VASC.AI against baseline ChatGPT models using 244 multiple-choice questions derived from the Vascular Education and Self-Assessment Program (VESAP-5). Building on these encouraging findings, we then created the **Artificial Intelligence – Assessment of Standards, Compliance, Efficacy, Navigation, & Decision-Making (AI-ASCEND) Benchmark** benchmark—a series of standardized clinical scenarios focusing on complex aortic pathologies—to compare VASC.AI with not only ChatGPT models but also with other general-purpose LLMs.

METHODS

Study Design Overview

Our research was conducted in two sequential phases. The first phase (VESAP Study) focused on evaluating educational performance using a bank of multiple-choice questions from the Vascular Education and Self-Assessment Program (VESAP-5). The second phase (ASCEND-AI Benchmark) aimed to assess clinical decision-making in complex aortic pathology scenarios. Both phases compared VASC.AI—a vascular surgery–specific LLM employing Retrieval-Augmented Generation (RAG) (**Figure 1**)—against general-purpose LLMs (including various versions of ChatGPT, Gemini, and Copilot).

Educational Assessment (VESAP Study)

Dataset and Question Selection: We curated 244 text-based multiple-choice questions from six VESAP-5 modules, covering aortoiliac, cerebrovascular, lower extremity, renal/mesenteric, vascular medicine, and venous disease. Questions were selected to represent a broad range of difficulty levels and content areas relevant to vascular surgery, ensuring that both fundamental and advanced clinical topics were addressed.

***Administration and Response Collection*:** Each question was input directly into the baseline models (ChatGPT-3.5, ChatGPT-4, and ChatGPT-4o) and into VASC.AI using a standardized prompt format to eliminate bias. To control for contextual drift, each question was entered in a new session ensuring responses were generated independently.

***Scoring and Error Classification*:** The primary outcome was the percentage of correct responses. Additionally, responses that were incorrect were further analyzed and classified into two error categories: Logical Errors - These errors occurred when the model identified relevant information but failed to properly integrate or apply it to arrive at the correct answer. Information Errors - These involved omissions where key details from the question were not recognized or retrieved, leading to an incomplete or incorrect response. Three independent evaluators (with vascular surgery expertise) assessed each response. Discrepancies were resolved via group consensus to ensure consistency in scoring.

Clinical Decision-Making Assessment (ASCEND-AI Benchmark)

Clinical Scenario Development: A set of 35 standardized clinical scenarios was developed, focusing on complex aortic pathologies, including thoracic and abdominal aneurysms, dissections, and occlusive disease. These scenarios were designed to mirror realistic patient presentations with multiple diagnostic possibilities and required comprehensive workup and management plans.

Evaluation Domains and Scoring Rubric: Responses from each platform (VASC.AI, ChatGPT-4, Gemini, and Copilot) were evaluated across five domains, Accuracy of Differential Diagnosis, Thoroughness of Workup Suggestions, Clarity in Medical Optimization, Relevance of Treatment Options, and Overall Usefulness in Decision-Making. Each domain was scored on a scale of 1 (poor) to 5 (excellent), with an overall composite score calculated for each response.

Data Collection and Blinding: Responses were collected in a standardized manner, with each clinical scenario administered separately to avoid context contamination. Scoring was performed by a panel of vascular surgery experts who were blinded to the identity of the responding model to minimize potential evaluator bias.

Statistical Analysis: Mean scores and standard deviations were computed for each evaluation domain as well as the overall composite score for each model. One-way ANOVA was utilized to assess significant differences in performance across the models. For domains showing statistically significant differences, pairwise t-tests were conducted to further delineate which models differed significantly from one another. A p-value of < 0.05 was set as the threshold for statistical significance. GraphPad Prism (version 9) was used for all statistical calculations and to generate corresponding figures (e.g., bar graphs, error bars) that clearly illustrate inter-model performance differences.

RESULTS

Educational Assessment (VESAP Study):

In the VESAP study, 244 multiple-choice questions from six VESAP-5 modules were used to evaluate performance. The baseline models yielded correct response rates as follows, **ChatGPT-3.5:** 55.3% (±4.3%), **ChatGPT-4:** 69.0% (±4.9%), **ChatGPT-4o:** 77.7% (±7.6%). In contrast, **VASC.AI** achieved a markedly higher correct response rate of 93.8% (±2.4%), a statistically significant improvement over all ChatGPT models (p < 0.001). Error analysis further highlighted the advantages of VASC.AI (**Figure 2**). The incorrect responses in ChatGPT-3.5 comprised roughly 34.5% logical errors and 65.5% information errors, indicating issues both in integrating retrieved data and in omitting critical facts. ChatGPT-4 and ChatGPT-4o exhibited similar mixed error profiles. Notably, all errors observed in VASC.AI were purely logical—there were no information errors, demonstrating that its RAG-enhanced framework effectively minimizes factual inaccuracies by grounding responses in a comprehensive, specialized database.

Clinical Decision-Making Assessment (ASCEND Benchmark):

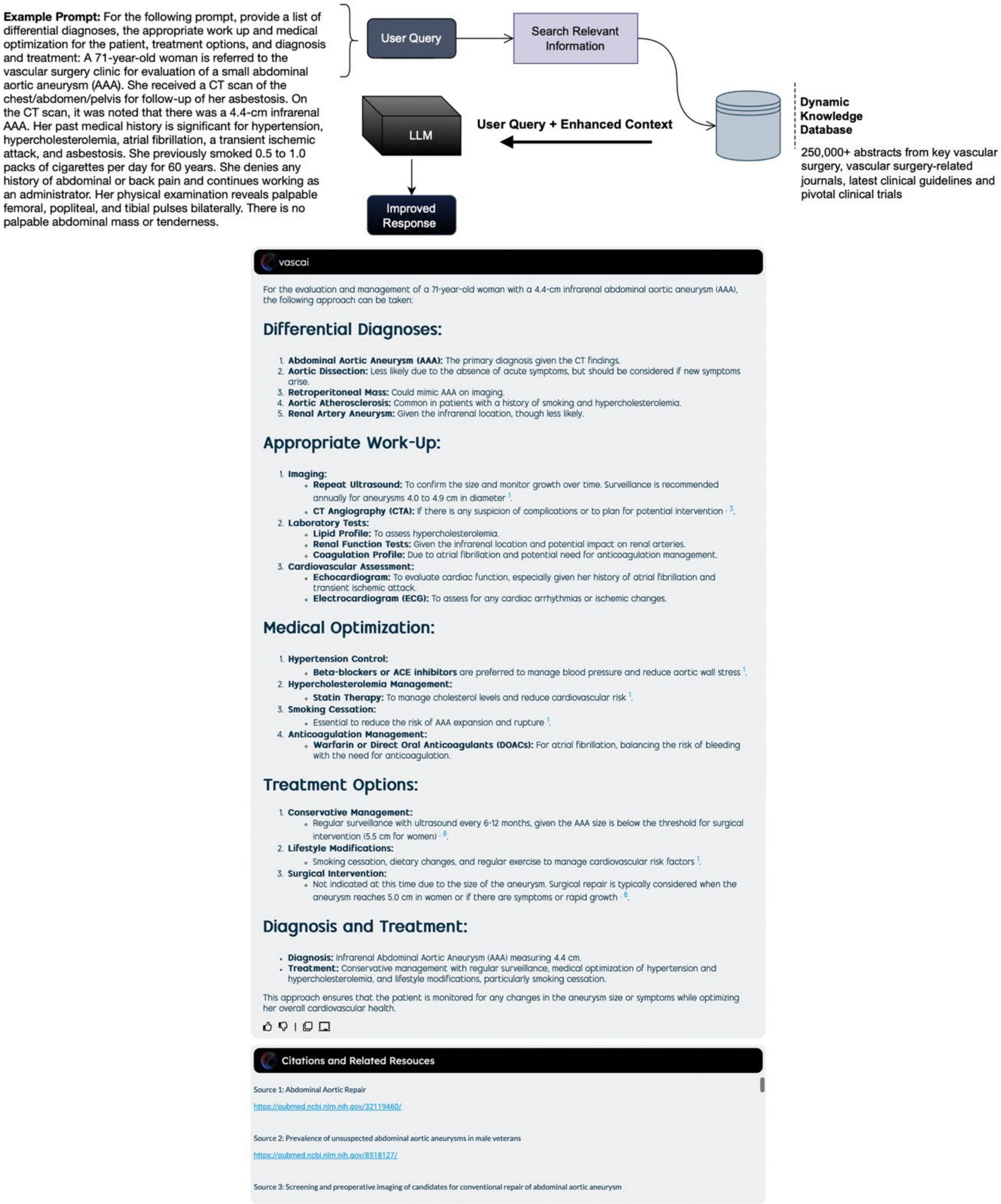
The AI-ASCEND benchmark employed 35 standardized clinical scenarios focused on complex aortic pathologies. Evaluation by vascular surgery experts across five key domains revealed that VASC.AI consistently outperformed its competitors. Specifically, in **Accuracy of Differential Diagnosis,** VASC.AI scored 4.80 ± 0.25. In comparison, ChatGPT-4 scored 4.60 ± 0.40, Gemini 4.10 ± 0.30, and Copilot 4.30 ± 0.35, indicating a more precise identification of potential diagnoses by VASC.AI**.** For **Thoroughness of Workup Suggestions**, VASC.AI attained a score of 4.90 ± 0.15, exceeding ChatGPT-4 (4.50 ± 0.30), Gemini (4.20 ± 0.20), and Copilot (4.10 ± 0.25). This reflects VASC.AI's ability to propose more comprehensive and prioritized diagnostic plans. Regarding **Clarity in Medical Optimization**, VASC.AI led with a score of 4.85 ± 0.20, compared to ChatGPT-4 at 4.40 ± 0.25, Gemini at 4.00 ± 0.30, and Copilot at 3.80 ± 0.35. This suggests that VASC.AI provided clearer, more actionable optimization strategies. For **Relevance of Treatment Options**, VASC.AI achieved the highest score at 4.95 ± 0.10, while ChatGPT-4, Gemini, and Copilot scored 4.50 ± 0.20, 4.10 ± 0.25, and 4.05 ± 0.30, respectively. Finally, for **Overall Usefulness in Decision-Making**, VASC.AI achieved a mean score of 4.90 ± 0.15, which was notably higher than the scores obtained by ChatGPT-4 (4.50 ± 0.25), Gemini (4.00 ± 0.30), and Copilot (4.00 ± 0.35) (**Figure 3**). A one-way ANOVA was conducted to assess overall performance differences among the four platforms, revealing a statistically significant variation ($p < 0.0001$). Pairwise t-tests further elucidated these differences, showing that VASC.AI significantly outperformed all other models. Specifically, VASC.AI showed superior performance compared with ChatGPT-4o ($t = 5.91$, $p < 0.0001$), Gemini ($t = 15.18$, $p < 0.0001$), and Copilot ($t = 15.00$, $p < 0.0001$). ChatGPT-4o also demonstrated significant advantages over both Gemini ($t = 6.49$, $p < 0.0001$) and Copilot ($t = 5.73$, $p < 0.0001$).

CONCLUSIONS

Our study demonstrates that a domain-specific AI platform, VASC.AI, which leverages Retrieval-Augmented Generation to integrate a curated, up-to-date vascular surgery knowledge base, significantly outperforms general-purpose LLMs in both educational and clinical decision-making settings. The markedly higher correct response rate in the VESAP study (93.8% versus 55.3–77.7% for ChatGPT models) underscores its potential to enhance medical education by delivering accurate, contextually relevant information with minimal factual errors. In addition, the AI-ASCEND Benchmark revealed that VASC.AI consistently achieved superior scores across critical clinical domains—ranging from differential diagnosis and workup thoroughness to clarity in medical optimization and treatment relevance. These findings collectively indicate that the specialized design of VASC.AI not only minimizes information errors but also translates into improved overall clinical utility.

The enhanced performance of VASC.AI can be attributed to its tailored integration of RAG technology, which grounds its outputs in a robust database of vascular-specific literature, guidelines, and clinical trials. This approach ensures that recommendations are evidence-based and finely tuned to the complexities of vascular surgery, thereby bridging the gap between theoretical knowledge and real-world clinical decision-making. The consistently higher scores across all evaluation domains suggest that such specialized AI platforms may have a significant role in advancing both patient care and surgical education. Furthermore, while our findings are highly promising, they also highlight areas for future improvement. For instance, although VASC.AI's errors were limited to logical missteps, further refinement is needed to enhance its reasoning processes. Future work should focus on integrating multimodal data, such as imaging and laboratory results, to provide a more comprehensive clinical evaluation. Additionally, expanding the knowledge base to include emerging research and a broader spectrum of vascular conditions could further solidify the role of VASC.AI as a transformative tool in precision medicine.

In conclusion, our research supports the adoption of specialized, RAG-enhanced AI platforms like VASC.AI in vascular surgery. By providing clinicians and trainees with highly accurate, context-aware recommendations, these systems promise to improve decision-making efficiency, elevate educational outcomes, and ultimately contribute to better patient care.



Example Prompt: For the following prompt, provide a list of differential diagnoses, the appropriate work up and medical optimization for the patient, treatment options, and diagnosis and treatment: A 71-year-old woman is referred to the vascular surgery clinic for evaluation of a small abdominal aortic aneurysm (AAA). She received a CT scan of the chest/abdomen/pelvis for follow-up of her asbestosis. On the CT scan, it was noted that there was a 4.4-cm infrarenal AAA. Her past medical history is significant for hypertension, hypercholesterolemia, atrial fibrillation, a transient ischemic attack, and asbestosis. She previously smoked 0.5 to 1.0 packs of cigarettes per day for 60 years. She denies any history of abdominal or back pain and continues working as an administrator. Her physical examination reveals palpable femoral, popliteal, and tibial pulses bilaterally. There is no palpable abdominal mass or tenderness.

User Query
Search Relevant Information
Dynamic Knowledge Database
250,000+ abstracts from key vascular surgery, vascular surgery-related journals, latest clinical guidelines and pivotal clinical trials
User Query + Enhanced Context
LLM
Improved Response
vascai

For the evaluation and management of a 71-year-old woman with a 4.4-cm infrarenal abdominal aortic aneurysm (AAA), the following approach can be taken:

Differential Diagnoses:

1. **Abdominal Aortic Aneurysm (AAA):** The primary diagnosis given the CT findings.
2. **Aortic Dissection:** Less likely due to the absence of acute symptoms, but should be considered if new symptoms arise.
3. **Retroperitoneal Mass:** Could mimic AAA on imaging.
4. **Aortic Atherosclerosis:** Common in patients with a history of smoking and hypercholesterolemia.
5. **Renal Artery Aneurysm:** Given the infrarenal location, though less likely.

Appropriate Work-Up:

1. **Imaging:**
   - **Repeat Ultrasound:** To confirm the size and monitor growth over time. Surveillance is recommended annually for aneurysms 4.0 to 4.9 cm in diameter [1].
   - **CT Angiography (CTA):** If there is any suspicion of complications or to plan for potential intervention [3].
2. **Laboratory Tests:**
   - **Lipid Profile:** To assess hypercholesterolemia.
   - **Renal Function Tests:** Given the infrarenal location and potential impact on renal arteries.
   - **Coagulation Profile:** Due to atrial fibrillation and potential need for anticoagulation management.
3. **Cardiovascular Assessment:**
   - **Echocardiogram:** To evaluate cardiac function, especially given her history of atrial fibrillation and transient ischemic attack.
   - **Electrocardiogram (ECG):** To assess for any cardiac arrhythmias or ischemic changes.

Medical Optimization:

1. **Hypertension Control:**
   - **Beta-blockers or ACE inhibitors** are preferred to manage blood pressure and reduce aortic wall stress [1].
2. **Hypercholesterolemia Management:**
   - **Statin Therapy:** To manage cholesterol levels and reduce cardiovascular risk [1].
3. **Smoking Cessation:**
   - Essential to reduce the risk of AAA expansion and rupture [1].
4. **Anticoagulation Management:**
   - **Warfarin or Direct Oral Anticoagulants (DOACs):** For atrial fibrillation, balancing the risk of bleeding with the need for anticoagulation.

Treatment Options:

1. **Conservative Management:**
   - Regular surveillance with ultrasound every 6-12 months, given the AAA size is below the threshold for surgical intervention (5.5 cm for women) [8].
2. **Lifestyle Modifications:**
   - Smoking cessation, dietary changes, and regular exercise to manage cardiovascular risk factors [1].
3. **Surgical Intervention:**
   - Not indicated at this time due to the size of the aneurysm. Surgical repair is typically considered when the aneurysm reaches 5.0 cm in women or if there are symptoms or rapid growth [8].

Diagnosis and Treatment:

- **Diagnosis:** Infrarenal Abdominal Aortic Aneurysm (AAA) measuring 4.4 cm.
- **Treatment:** Conservative management with regular surveillance, medical optimization of hypertension and hypercholesterolemia, and lifestyle modifications, particularly smoking cessation.

This approach ensures that the patient is monitored for any changes in the aneurysm size or symptoms while optimizing her overall cardiovascular health.

Citations and Related Resouces

Source 1: Abdominal Aortic Repair

https://pubmed.ncbi.nlm.nih.gov/32119460/

Source 2: Prevalence of unsuspected abdominal aortic aneurysms in male veterans

https://pubmed.ncbi.nlm.nih.gov/8518127/

Source 3: Screening and preoperative imaging of candidates for conventional repair of abdominal aortic aneurysm

Figure 1. VASC.AI RAG LLM Workflow and Example Response – An illustration of the VASC.AI Retrieval-Augmented Generation (RAG) LLM Model's process for diagnosing and managing complex aortic pathologies, accompanied by an example response to a clinical scenario involving an asymptomatic 4.4 cm infrarenal abdominal aortic aneurysm.

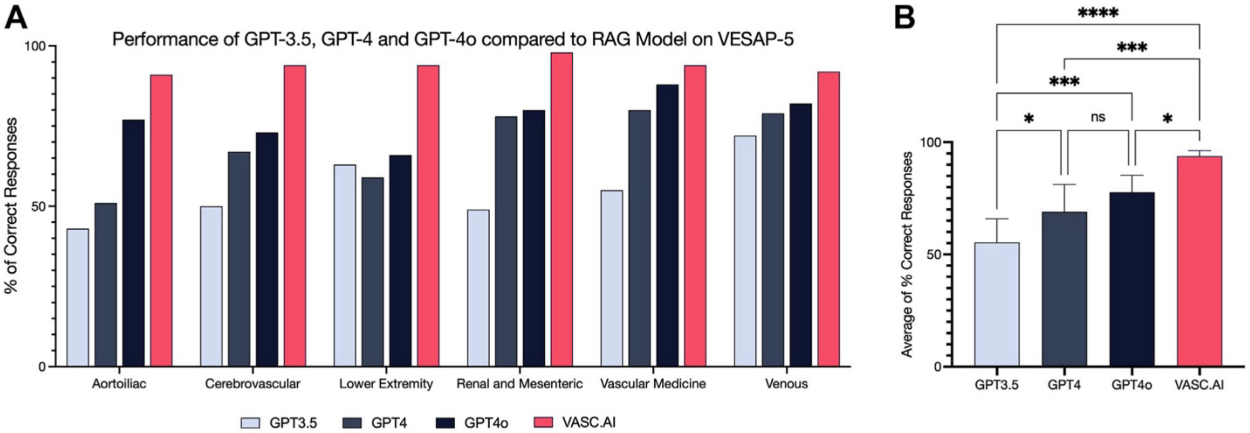


Figure 2. Performance of ChatGPT-3.5, ChatGPT-4, and ChatGPT-4o compared with retrieval-augmented generation (RAG) large language model (LLM), VASC.AI on Vascular Education and Self-assessment Program (VESAP) 5 modules. (A) Percentage of correct responses from RAG LLM, ChatGPT 3.5, 4, and 4o for each VESAP section, with the **(B)** comparison of the average of overall correct responses. Values are means 6 standard deviations. Data were analyzed using two-way analysis of variance (P < .05), Tukey's multiple comparison test. *P < .01; **P < .01; ***P < .001; ****P < .0001; ns, not significant.

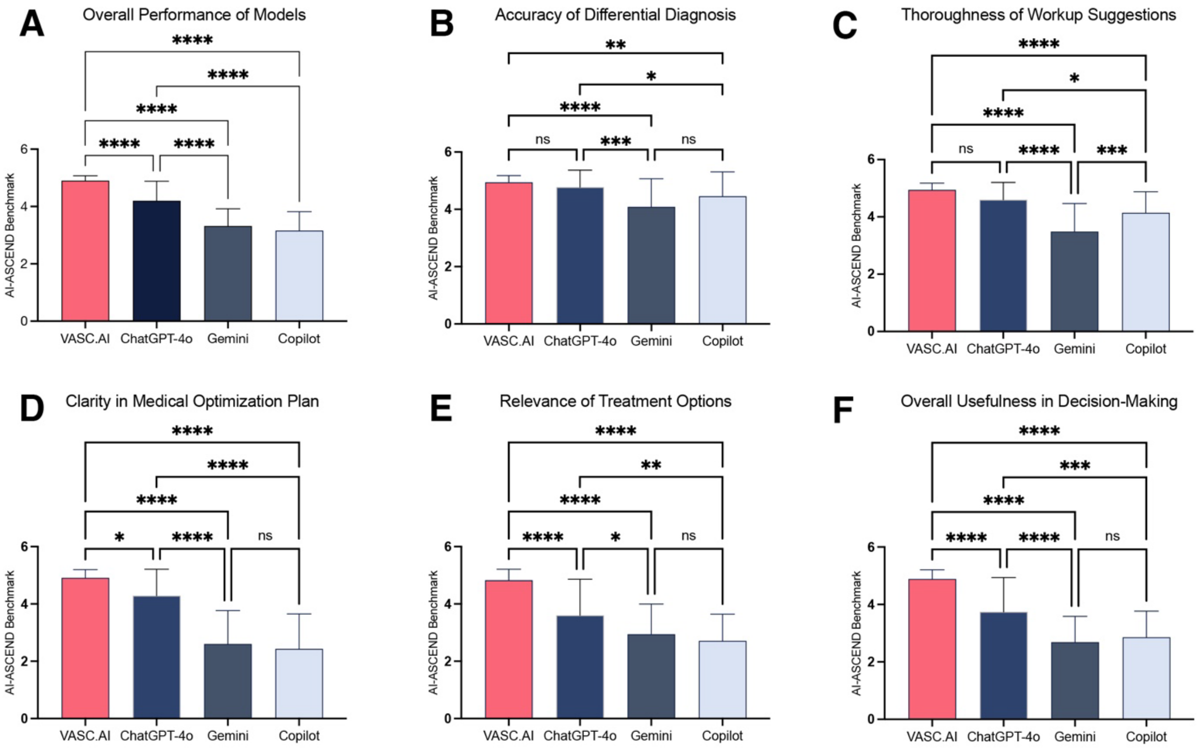


Figure 3. Overall and Section-Specific Performance of AI Platforms. Overall and section-specific performance of VASC.AI, ChatGPT-4o, Gemini, and Copilot in clinical scenarios involving complex aortic pathologies, as assessed by the AI-ASCEND Benchmark. Subpanels depict **(A)** Overall Performance of Models, **(B)** Accuracy of Differential Diagnosis, **(C)** Thoroughness of Workup Suggestions, **(D)** Clarity in Medical Optimization Plan, **(E)** Relevance of Treatment Options, and **(F)** Overall Usefulness in Decision-Making. Values are presented as means ± standard deviation. Data were analyzed using one-way analysis of variance (ANOVA) followed by post hoc pairwise t-tests. *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001; ns = not significant.

PRIMARY CARE ACCESS IS ASSOCIATED WITH IMPROVED LONG-TERM SURVIVAL AFTER SEVERE TRAUMATIC INJURY

Elliott K. Yee (SSTP)[1], Stephanie A. Mason[1], Laura C. Rosella[2], Liisa Jaakkimainen[3], Brandon M. Zagorski[4], Darby Little[5], Gemma Postill[6], Avery B. Nathens[1], Bourke W. Tillmann[7], Barbara Haas[1,7]

[1]Division of General Surgery, Department of Surgery, University of Toronto, Toronto, Ontario
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario
[3]Department of Family and Community Medicine, University of Toronto, Toronto, Ontario
[4]ICES
[5]Division of Plastic, Reconstructive & Aesthetic Surgery, Department of Surgery, University of Toronto, Toronto, Ontario
[6]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario
[7]Interdepartmental Division of Critical Care, University of Toronto, Toronto, Ontario

The authors have decided not to make the research results available at this time and will provide updates as soon as the results can be shared.

Lomitapide enhances cytotoxic effects of temozolomide in chemo-resistant glioblastoma

Alyona Ivanova[1,2], Taylor Wilson[1,2], Kimia Ghannad-Zadeh[1,2], Esmond Tse[1], Megan Wu[1], Sunit Das[1-4]

[1]The Arthur and Sonia Labatt Brain Tumor Research Center, The Hospital for Sick Children, Toronto, Canada

[2] Institute of Medical Sciences, University of Toronto, Toronto, Canada

[3] Division of Neurosurgery, University of Toronto, Toronto, Canada

[4] Keenan Chair in Surgery, St. Michael's Hospital, University of Toronto, Toronto, Canada

INTRODUCTION

Glioblastoma (GBM) is the most prevalent and aggressive malignant primary brain tumour in adults, with a median survival following multi-modality therapy of 14.6 months.[1] The current standard of care for patients with glioblastoma includes maximal safe surgical resection followed by radiation and chemotherapy with the alkylating agent, temozolomide (TMZ).[2] However, more than one-third of patients experience tumour progression during conventional therapy,[1] suggesting that many of these patients harbor tumour cells that are intrinsically resistant to temozolomide-associated cytotoxicity. There is significant need for novel therapies to complement or improve our current treatments.

METHODS

Study Design

Drug repurposing has gained attention in cancer research for its time- and monetary-efficiency in advancing chemical leads for clinical studies.[3] The use of FDA-approved agents significantly decreases the time required for agents to go from bench to bedside, as toxicological data for these drugs is publicly available.[4]

In this study, we performed a high-throughput drug screen using a library of approximately 900 FDA-approved candidates capable of crossing the blood-brain barrier (BBB). We identified eight agents that were predicted to have good brain penetration and exhibited cytotoxicity against glioma cells when combined with TMZ. As multiple recent reports have identified cholesterol biosynthesis as a vulnerability in glioblastoma,[5] we chose to investigate the lipid-lowering drug (statin), lomitapide (Juxtapid).

RESULTS

Lomitapide exerts cytotoxic effects on glioma cells and sensitizes them to the effects of temozolomide

To assess the drug response of lomitapide in glioma cells, a dose-response viability curve was constructed for CTL- and TR-U251 (treatment-resistant) cell lines treated with lomitapide for 72 hours. Lomitapide produced characteristic dose-response relationship in an exploitable dosage range and showed to be beneficial for GBM chemo-resistant therapy. Lomitapide treatment resulted in cell death in both CTL-U251 and TR-U251 cell lines in a dose-dependent manner, with minimal effect on viability of normal HEK293 cells (Figure 1A).

Concomitant treatment with lomitapide and TMZ resulted in a statistically significant decrease in viability in both U251 and TR-U251 cells, compared to treatment with TMZ alone (Figure 1B-C). Notably, when TMZ was combined with lomitapide, 1/10 times lower TMZ dosage (10μM) induced just as much cytotoxicity in CTL-U251 and TR-U251 cells as 100μM TMZ combined with 2μM lomitapide (Figure 1I-J). GBM patients are therefore provided with a new therapeutic potential that minimizes chemotoxicity, increases overall quality of life and eventually improves patient outcome.

As glioma stem cells (GSCs) have been postulated to be the drivers of tumour progression and treatment resistance, we next examined the effect of lomitapide in multiple GSC cell lines. Briefly, three GSC lines (GliNS1, 811, 818) were treated with lomitapide alone, TMZ alone, or concomitant lomitapide and TMZ. Concomitant treatment with lomitapide and TMZ resulted in a statistically significant decrease in cell viability in all three GSC lines, compared to treatment with TMZ alone (Figure 1D-F).

Lomitapide inhibits *de novo* cholesterol synthesis

The use of statins (lipid-lowering drugs) as anti-neoplastic agents has recently gained attention in oncology.[7] Statins utilize the dependency of highly proliferative tumour cells on the mevalonate pathway - an essential biosynthetic step that provides the precursors for *de novo* synthesis of cholesterol.[7] Statins block mevalonate synthesis by inhibiting the rate-limiting step of this pathway.[7] We hypothesized that the effect of lomitapide on glioma cells was due to the statin-like mechanism of function as an inhibitor of the mevalonate pathway (Figure 2B). Inhibition of the mevalonate pathway by lomitapide in CTL-U251 and TR-U251 cells was verified by the accumulation of the rate-limiting enzyme, 3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR) (Figure 2C-D, F-G). HMGCR did not accumulate in lomitapide-treated HEK293 cells (Figure 2E), emphasizing that lomitapide is selective towards neoplastic cells with increased mevalonate pathway flux that supports oncogenesis.

Notably, lomitapide treatment resulted in a significant decrease in cholesterol uptake by glioma cells cultured in serum-free media (Figure 2H), in a dose-dependent manner. Since we did not observe any change in expression of lipid transfer protein MTTP(Figure 2A), we postulate that decreased cholesterol uptake in glioma cells with lomitapide treatment is due to depletion of cholesterol in the media secondary to inhibition of intracellular cholesterol synthesis.

Lomitapide primes glioma cells for ferroptosis vulnerability via depletion of CoQ_{10} and production of ROS

In addition to its role in cholesterol synthesis, the mevalonate pathway is also critical for production of the mitochondrial protein, coenzyme Q_{10} (CoQ_{10}).[8] CoQ_{10} is a universal multifunctional lipid with fundamental roles in cellular homeostasis. Besides its metabolic and bioenergetic functions in mitochondria, CoQ_{10} acts as a lipophilic antioxidant and supports enzymatic reactions in other cellular membranes.[9]

We hypothesized that mevalonate pathway inhibition by lomitapide would deplete CoQ_{10} levels in highly proliferative glioma cells that rely on mevalonate pathway and cause oxidative stress, priming them for chemotherapy-induced cell death (Figure 2I). Indeed, treatment with lomitapide resulted in CoQ_{10} depletion in both CTL-U251 and TR-U251 cells (Figure 2J). Subsequently, cellular ROS were significantly elevated in lomitapide-treated CTL-U251 and TR-U251 cells, but not in HEK293 cells (Figure 2K). These findings indicate that while lomitapide has a pronounced effect in glioma cells, it exerts minimum cytotoxicity in normal cells.

Our *in vitro* findings support the conclusion that lomitapide enhances glioma cell sensitivity to TMZ through deregulation of mevalonate pathway and induction of ferroptosis.

Lomitapide delays tumour recurrence and improves survival when combined with TMZ in a glioblastoma xenograft mouse model

We evaluated the effect of concurrent lomitapide treatment with TMZ on tumour growth and survival using a mouse xenograft model. Following intracranial inoculation with luciferase-expressing U251 cells, mice were allocated into four treatment groups: 1) control (no treatment); 2) 3 cycles of lomitapide (7.6 mg/kg/day) alone; 3) TMZ (10 mg/kg/day) alone for a 1-week; 4) combination therapy with lomitapide and TMZ (Figure 3A).

Control and lomitapide alone-treated mice displayed weekly advances in tumour growth and all succumbed to their disease within 4 weeks of implantation. While lomitapide alone demonstrated little to no efficacy on survival or anti-tumour activity at the concentration used for treatment, (Figures 3B and 3C), the addition of lomitapide to TMZ therapy resulted in a significant reduction in tumour growth (Figures 3B-D) and prolongation in mouse survival (Figure 3E), compared to treatment with TMZ alone (Figure 3B-E). Tumour progression was significantly delayed in mice receiving combination therapy (Figure 3C), compared to those treated with TMZ alone.

CONCLUSIONS

Our findings identify lomitapide as a potential therapeutic agent capable of targeting treatment resistance and delaying tumour progression in glioblastoma.

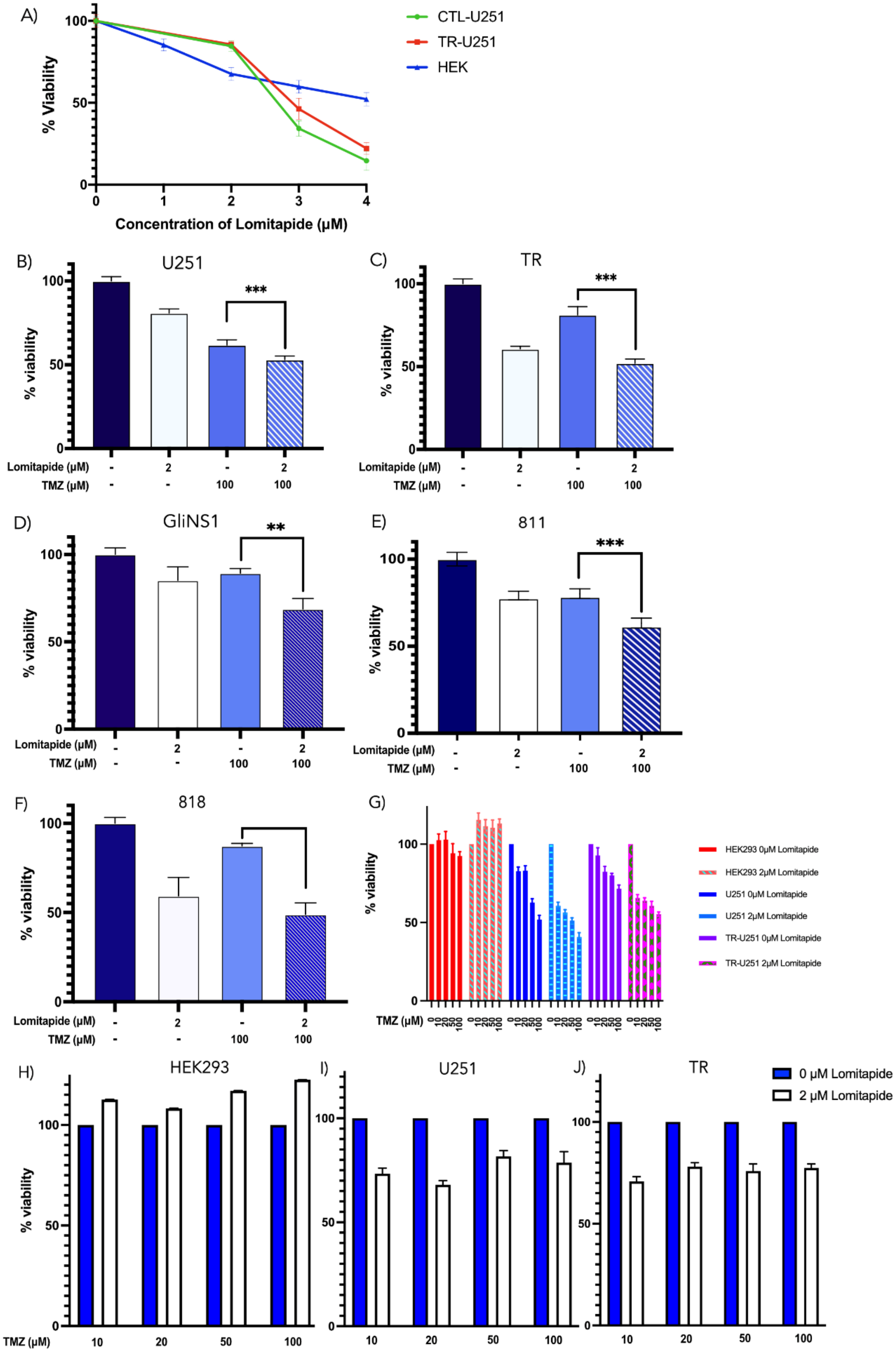

A)
% Viability
100
50
0
0
1
2
3
4
Concentration of Lomitapide (μM)
CTL-U251
TR-U251
HEK
B)
U251
% viability
100
50
0
***
Lomitapide (μM) - 2 - 2
TMZ (μM) - - 100 100
C)
TR
% viability
100
50
0
***
Lomitapide (μM) - 2 - 2
TMZ (μM) - - 100 100
D)
GliNS1
% viability
100
50
0
**
Lomitapide (μM) - 2 - 2
TMZ (μM) - - 100 100
E)
811
% viability
100
50
0
***
Lomitapide (μM) - 2 - 2
TMZ (μM) - - 100 100
F)
818
% viability
100
50
0
Lomitapide (μM) - 2 - 2
TMZ (μM) - - 100 100
G)
% viability
100
50
0
TMZ (μM) 0 10 20 50 100 0 10 20 50 100 0 10 20 50 100 0 10 20 50 100 0 10 20 50 100 0 10 20 50 100
HEK293 0μM Lomitapide
HEK293 2μM Lomitapide
U251 0μM Lomitapide
U251 2μM Lomitapide
TR-U251 0μM Lomitapide
TR-U251 2μM Lomitapide
H)
HEK293
% viability
100
50
0
TMZ (μM) 10 20 50 100
I)
U251
100
50
0
10 20 50 100
J)
TR
100
50
0
10 20 50 100
0 μM Lomitapide
2 μM Lomitapide

Figure 1. Lomitapide shows potential to target GBM cells and sensitize them to the effects of temozolomide, while minimizing damage to normal cells *in vitro*.

Cells were treated with TMZ with or without 2μM lomitapide. Cytotoxicity was estimated at 72 hrs post-treatment by Cell Viability Assay. **(A)** Viability assay shows CTL-U251 (green), TR-U251 (red) and HEK-293 (blue) cell response after treatment with serially increasing concentrations of lomitapide. Compared to TMZ alone, CTL-U251 **(B)** and TR-U251 **(C)** cell lines treated concomitantly with lomitapide (2 μM) and TMZ (100 μM) significantly decrease cell viability ($p<0.001$). Cell viability analysis on GBM stem cell lines GliNS1 cells **(D)**, 811 cells **(E)**, 818 cells **(F)** demonstrates that a concentration of lomitapide near its IC80 value (2 μM) added in combination with TMZ (100 μM) results in a significant reduction in cell viability, as compared to TMZ treatment alone ($p<0.01$, $p<0.001$, $p<0.001$, respectively). **(G)** Normal HEK293 cells, CTL-U251 and TR-U251 cells were treated with various doses of TMZ ranging from 0 to 100μM with or without 2μM lomitapide. Cell viability was normalized to untreated control. HEK293 cells **(H)**, CTL-U251 cells **(I)** and TR-U251 cells **(J)** were treated with TMZ combined with 2μM of lomitapide. Cell viability was normalized to TMZ treatment without lomitapide.

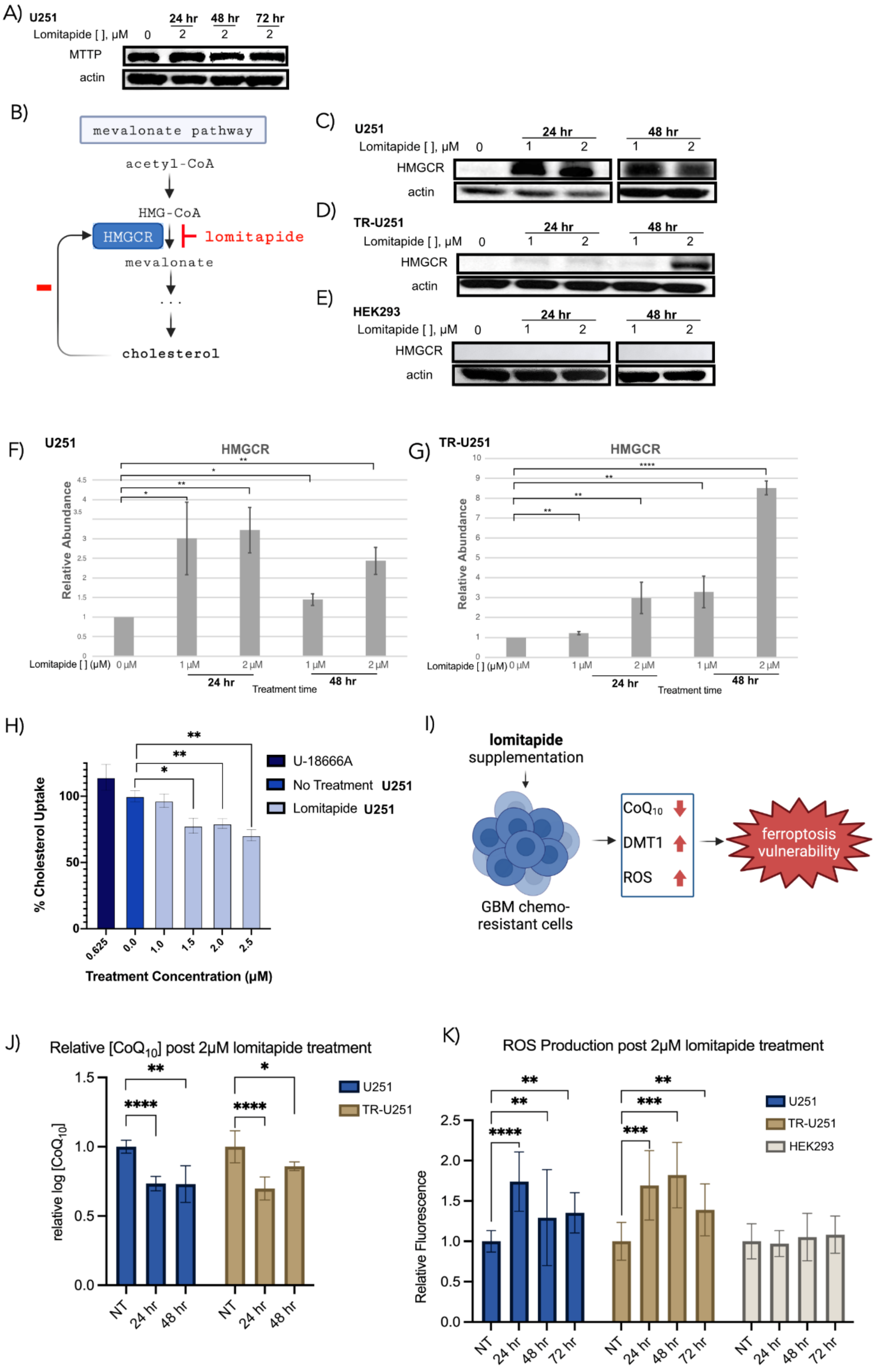

A)
U251
Lomitapide [ ], μM
0
24 hr
2
48 hr
2
72 hr
2
MTTP
actin
B)
mevalonate pathway
acetyl-CoA
HMG-CoA
HMGCR
lomitapide
mevalonate
. . .
cholesterol
C)
U251
Lomitapide [ ], μM
0
24 hr
1
2
48 hr
1
2
HMGCR
actin
D)
TR-U251
Lomitapide [ ], μM
0
24 hr
1
2
48 hr
1
2
HMGCR
actin
E)
HEK293
Lomitapide [ ], μM
0
24 hr
1
2
48 hr
1
2
HMGCR
actin
F)
U251
HMGCR
Relative Abundance
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
*
**
*
**
Lomitapide [ ] (μM)
0 μM
1 μM
2 μM
1 μM
2 μM
24 hr
Treatment time
48 hr
G)
TR-U251
HMGCR
Relative Abundance
0
1
2
3
4
5
6
7
8
9
10
**
**
**
****
Lomitapide [ ] (μM)
0 μM
1 μM
2 μM
1 μM
2 μM
24 hr
Treatment time
48 hr
H)
% Cholesterol Uptake
0
50
100
**
**
*
U-18666A
No Treatment U251
Lomitapide U251
0.625
0.0
1.0
1.5
2.0
2.5
Treatment Concentration (μM)
I)
lomitapide
supplementation
GBM chemo-
resistant cells
CoQ10
DMT1
ROS
ferroptosis
vulnerability
J)
Relative [CoQ10] post 2μM lomitapide treatment
relative log [CoQ10]
0.0
0.5
1.0
1.5
**
****
*
****
U251
TR-U251
NT
24 hr
48 hr
NT
24 hr
48 hr
K)
ROS Production post 2μM lomitapide treatment
Relative Fluorescence
0.0
0.5
1.0
1.5
2.0
2.5
**
**
****
**
***
***
U251
TR-U251
HEK293
NT
24 hr
48 hr
72 hr
NT
24 hr
48 hr
72 hr
NT
24 hr
48 hr
72 hr

Figure 2. Lomitapide cytotoxicity *in vitro* is attributed to inhibition of *de novo* cholesterol synthesis which primes glioma cells for ferroptosis via depletion of CoQ_{10} and production of ROS

(A) Lomitapide (2µM) treatment has no effect on the expression of microsomal triglyceride transport protein (MTTP) in CTL-U251 cells as confirmed by Western blot analysis post 24-, 48- and 72 hours of drug treatment. **(B)** Proposed mechanism of action of lomitapide. Lomitapide inhibits the critical rate-limiting step of mevalonate pathway by binding to HMGCR. While normally cholesterol production acts through a negative feedback loop to control HMGCR synthesis, reduction in cholesterol production causes accumulation of HMGCR. Inhibition of mevalonate pathway in CTL-U251 (**C, F**) and TR-U251 (**D, G**) cells at 24 and 48 hours following lomitapide treatment (0 µM, 1 µM, or 2 µM) is verified by accumulation of HMGCR. Lomitapide treatment does not lead to accumulation of HMGCR in normal HEK293 cells (**E**). **(H)** Cholesterol uptake by CTL-U251 cells following a 72-hour incubation in serum-free media at various lomitapide concentrations. U-18666A represents positive uptake control. **(I)** Mevalonate pathway inhibition depletes CoQ_{10} levels, results in ROS accumulation, which primes the cells for ferroptosis and makes them more susceptible to TMZ treatment. **(J)** Relative CoQ10 concentration in CTL-U251 and TR-U251 cells following 24- or 48-hour lomitapide treatment (2 µM). **(K)** Cellular ROS production in CTL-U251, TR-U251 and HEK293 cells treated with 2 µM lomitapide for 24, 48 and 72 hours. (* $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$)

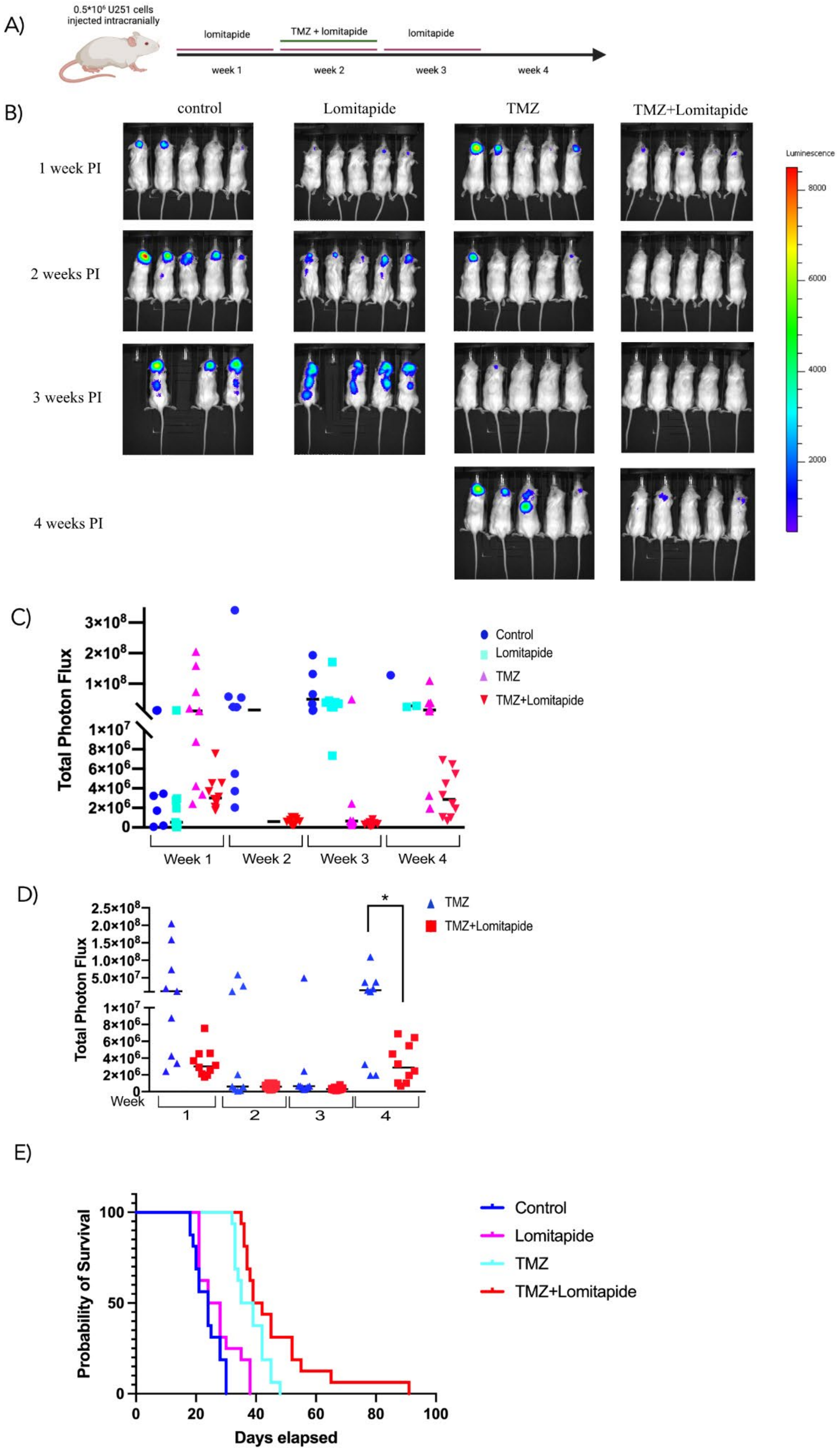

A)
0.5*10⁶ U251 cells injected intracranially
lomitapide
TMZ + lomitapide
lomitapide
week 1
week 2
week 3
week 4
B)
control
Lomitapide
TMZ
TMZ+Lomitapide
1 week PI
2 weeks PI
3 weeks PI
4 weeks PI
Luminescence
8000
6000
4000
2000
C)
Total Photon Flux
3×10⁸
2×10⁸
1×10⁸
1×10⁷
8×10⁶
6×10⁶
4×10⁶
2×10⁶
0
Week 1
Week 2
Week 3
Week 4
Control
Lomitapide
TMZ
TMZ+Lomitapide
D)
Total Photon Flux
2.5×10⁸
2.0×10⁸
1.5×10⁸
1.0×10⁸
5.0×10⁷
1×10⁷
8×10⁶
6×10⁶
4×10⁶
2×10⁶
0
Week
1
2
3
4
*
TMZ
TMZ+Lomitapide
E)
Probability of Survival
100
50
0
0
20
40
60
80
100
Days elapsed
Control
Lomitapide
TMZ
TMZ+Lomitapide

Figure 3. Lomitapide delays tumor recurrence and improves survival when combined with TMZ treatment in a glioblastoma xenograft model.

Post intracranial inoculation with luciferase-labeled U251 cells, mice were randomized into 4 groups. Bioluminescent images were recorded using an IVIS Lumina II Bioluminescence System (PerkinElmer) every 7 days. Total photon flux values were quantified in tumor progression between treatment groups as described in Sachdeva et al 2019.[6] **(A)** Schematic showing *in vivo* experimental design for U251 cells injected intracranially and thereafter treated with lomitapide, temozolomide (TMZ), or concomitant lomitapide and TMZ. **(B)** Bioluminescence imaging of intracranial glioblastoma mouse xenograft visualizing tumor growth of U251 cells. Empty spaces indicate sacrificed mice at humane endpoints. **(C)** Signal progression of total flux activity comparing tumor growth at 1- to 4-weeks post-inoculation. **(D)** Total photon flux of concomitant TMZ and lomitapide-treated and TMZ alone–treated mice 1-4 weeks following transplantation. **(E)** Kaplan–Meier survival curves of mice in individual cohorts (n=16). Median survival and statistical significance were determined by log-rank test: * p=0.0239 for lomitapide + TMZ vs. TMZ; ** p=0.0015 for lomitapide + TMZ vs. lomitapide; ** p=0.013 for lomitapide + TMZ vs. Control.

References

1. Ding, X., Zhang, W., Li, S., & Yang, H. (2019). The role of cholesterol metabolism in cancer. American Journal of Cancer Research, 9(2), 219-227.
2. Stupp, R., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine, 352, 987-996.
3. Abbruzzese, C., Matteoni, S., Signore, M., Cardone, L., Nath, K., Glickson, J. D., & Paggi, M. G. (2017). Drug repurposing for the treatment of glioblastoma multiforme. Journal of Experimental & Clinical Cancer Research, 36(1), 169. doi: 10.1186/s13046-017-0642-x.
4. Verbaanderd, C., Meheus, L., Huys, I., & Pantziarka, P. (2017). Repurposing Drugs in Oncology: Next Steps. Trends in Cancer, 3(8), 543-6. doi: 10.1016/j.trecan.2017.06.007.
5. Zhang, C., Liu, X., Jin, S., Chen, Y., & Guo, R. (2022). Ferroptosis in cancer therapy: a novel approach to reversing drug resistance. Molecular Cancer, 21, 47. doi: 10.1186/s12943-022-01530-y.
6. Sachdeva, R., Wu, M., Smiljanic, S., Kaskun, O., Ghannad-Zadeh, K., Celebre, A., Isaev, K., Morrissy, A. S., Guan, J., Tong, J., et al. (2019). ID1 is critical for tumourigenesis and regulates chemoresistance in glioblastoma. Cancer Research, 79(16), 4057-4071. doi: 10.1158/0008-5472.CAN-18-1357.
7. Juarez D, Fruman DA. (2021). Targeting the Mevalonate Pathway in Cancer. Trends in Cancer, 7(6), 525-540. doi: 10.1016/j.trecan.2020.11.008.
8. Stockwell, B. R. (2019). A powerful cell-protection system prevents cell death by ferroptosis. Nature, 575(7784), 597-598. doi: 10.1038/d41586-019-03145-8.
9. Saini, R. (2011). Coenzyme Q10: The essential nutrient. Journal of Pharmacy & Bioallied Sciences, 3(3), 466-467. doi: 10.4103/0975-7406.84471.

AN ARTIFICIAL INTELLIGENCE-BASED MODEL TO PREDICT PROGRESSION RISK IN NON-MUSCLE INVASIVE BLADDER CANCER (PROGRXN-BCA) AND IMPROVE SUBSTRATIFICATION OF INTERMEDIATE- AND HIGH-RISK PATIENTS: AN INTERNATIONAL MODEL DEVELOPMENT AND EVALUATION STUDY

Jethro C.C. Kwong (SSTP)[1,2,3], Zizo Al-Daqqaq[4], Yashan Chelliahpillai[4], Soomin Lee[4], Kellie Kim[4], Maximiliano Ringa[5], Andrew Feifer[1,2,5], Katherine Lajkosz[2], Marian S. Wettstein[1,2], Amy Chan[6], Taeweon Lee[7], Myky Nguyen[8], Wassim Kassouf[9], Peter C. Black[10], Rodney H. Breau[11], Michele Lodde[12], Adrian Fairey[13], Jean-Baptiste Lattouf[14], Claudio Jeldres[15], Ricardo Rendon[16], Nimira Alimohamed[17], Neil E. Fleshner[1,2], Romain Diamand[8], Paolo Gontero[18], Richard J. Sylvester[19], Bas W.G. van Rhijn[20], Ashish M. Kamat[7], Alistair E.W. Johnson[3,21,22], Alexandre R. Zlotta[1,2,6], Girish S. Kulkarni[1,2,3] on behalf of the PROGRxN-BCa consortium

[1]Division of Urology, Department of Surgery, University of Toronto, Toronto, Canada
[2]Division of Urology, Department of Surgery, University Health Network, Toronto, Canada
[3]Temerty Centre for AI Research and Education in Medicine, University of Toronto, Toronto, Canada
[4]Temerty Faculty of Medicine, University of Toronto, Toronto, Canada
[5]Division of Urology, Department of Surgery, Trillium Health Partners, Mississauga, Canada
[6]Division of Urology, Department of Surgery, Mount Sinai Hospital, Sinai Health System, Toronto, Canada
[7]Department of Urology, University of Texas MD Anderson Cancer Center, Houston, United States
[8]Department of Urology, Jules Bordet Institute-Erasme Hospital, Hôpital Universitaire de Bruxelles, Université Libre de Bruxelles, Brussels, Belgium
[9]Department of Urology, McGill University Health Centre, Montreal, Canada
[10]Department of Urologic Sciences, University of British Columbia, Vancouver, Canada
[11]Division of Urology, Department of Surgery, The Ottawa Hospital Research Institute, Ottawa, Canada
[12]Division of Urology, Department of Surgery, CHU de Québec-Université Laval, Quebec City, Canada
[13]Division of Urology, Department of Surgery, University of Alberta, Edmonton, Canada
[14]Division of Urology, Department of Surgery, Centre Hospitalier de l'Université de Montréal, Montreal, Canada
[15]Division of Urology, Department of Surgery, Université de Sherbrooke, Sherbrooke, Canada
[16]Department of Urology, Dalhousie University, Halifax, Canada
[17]Department of Oncology, Cumming School of Medicine, University of Calgary, Calgary, Canada
[18]Department of Urology, Città della Salute e della Scienza, University of Torino School of Medicine, Torino, Italy
[19]Department of Biostatistics, European Organization for Research and Treatment of Cancer, Brussels, Belgium
[20]Department of Surgical Oncology (Urology), Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands
[21]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[22]Vector Institute, Toronto, Canada

INTRODUCTION

Non-muscle invasive bladder cancer (NMIBC) is one of the most expensive malignancies to treat, due to the need for long-term surveillance and repeated transurethral resection of bladder tumours (TURBT).[1] Despite complete resection and adjuvant treatments, progression rates to potentially lethal muscle-invasive disease (MIBC) remain high and the window of cure is narrow, particularly for intermediate- and high-risk NMIBC.[2–4] Therefore, accurate prediction of tumour progression is essential for patient counselling, to guide timely treatment intensification, and to identify clinical trial candidates.

Risk calculators have historically been used to identify NMIBC patients at increased risk of progression.[5,6] However, their applicability to contemporary NMIBC patients is limited due to changes in management and their reliance on the World Health Organization (WHO) 1973 grading system, which is prone to substantial interobserver variability.[7] Current approaches to risk stratification are based on the European Association of Urology (EAU) NMIBC risk calculator, which is the only model that also incorporates the current WHO 2004/2022 grading system and classifies patients into low-, intermediate-, high-, or very high-risk groups.[8] However, patients treated with bacillus Calmette-Guérin (BCG) were excluded, which is now standard of care for intermediate- and high-risk NMIBC. On external validation, the EAU model has a c-index of 0.63.[9]

Artificial intelligence (AI) has shown promise in improving personalized prognostication and treatment in uro-oncology.[10] However, current evidence supporting AI applications in NMIBC is weak. Using APPRAISE-AI, a novel tool we developed to assess the methodological and reporting quality of AI studies, we found that most AI studies in NMIBC prognostication were low quality.[1,11] Common shortcomings included dataset limitations, inconsistent outcome definitions, methodological concerns, inadequate model evaluation, and reproducibility issues.

To address these limitations, we aimed to develop and validate PROGRxN-BCa (PROGression Risk assessment in NMIBC), a prognostic model to estimate the five-year risk of progression in NMIBC patients using the largest, international cohort of almost 13000 patients from both academic and community hospitals. A secondary aim was to ensure that PROGRxN-BCa was generalizable regardless of resource constraints, BCG shortages, or guideline adherence. Finally, we aimed to improve substratification of the highly heterogeneous intermediate- and high-risk groups to provide more precise risk estimates of disease progression.[12]

METHODS

Study design

This study was a supervised time-to-event analysis to predict time to progression using available information after TURBT. A total of 1956 patients with 382 progression events were required to satisfy the minimum sample size criteria of Riley et al.[13] This study was conducted following the STREAM-URO framework, a standardized reporting framework we previously developed for AI studies in urology.[14] Two data scientists independently analyzed the data to verify the results.

Data sources and eligibility criteria

The training cohort included patients who underwent TURBT at four Canadian academic and community-based hospitals between Jan 1, 2005 and Jun 30, 2022. An international, external testing cohort comprising of patients treated at 30 North American and European institutions between Jan 1, 2005 and Dec 31, 2023 was used (**Figure 1**). All NMIBC patients (Ta, T1, or primary CIS) were included regardless of prior tumour history. Patients received BCG or other intravesical chemotherapy at their physician's discretion. Exclusion criteria included muscle invasive disease (stage ≥ T2) at diagnosis, treatment before Jan 1, 2005, immediate cystectomy for NMIBC, benign pathology, and missing pathological information.

Data abstraction, processing, and outcome definition

Candidate features (variables) known before or at index TURBT were selected based on prior literature.[2,3] Missing data were imputed using HyperImpute (version 0.1.17), a generalized iterative imputation framework.[15] No other feature engineering or removal steps were performed.

Primary outcome was time to progression, defined as the interval from date of TURBT to date of first development of muscle-invasive (stage ≥ pT2 either at subsequent TURBT or cystectomy), nodal, and/or metastatic disease. This definition was chosen due to its treatment implications and its alignment with the EAU risk calculator. Patients without progression were censored at the date of last follow-up cystoscopy or date of death, whichever event occurred first.

PROGRxN-BCa development and explanations

PROGRxN-BCa was developed using a random survival forest (scikit-survival version 0.22.2). All candidate features were incorporated into the model. The final PROGRxN-BCa model was retrained on the full training cohort using the optimal set of hyperparameters. Feature importance was estimated using permutation importance (scikit-learn version 1.3.2), which measures the change in model performance when the relationship between a given feature and progression risk is disrupted by random shuffling.

Model evaluation

PROGRxN-BCa was compared to the widely-used EAU risk calculator, which is endorsed by the EAU NMIBC guidelines and is the only model that can incorporate either the WHO 1973 or the current 2004/2022 grading system.[8]

Model evaluation was based on discrimination, calibration, and net benefit. Five-year progression risk was used given its clinical relevance in guiding management and its availability as a prediction timepoint for the EAU risk calculator. Discrimination was assessed by concordance index (c-index). Smoothed calibration curves compared the predicted and observed risks of progression. Decision curve analysis measured the net benefit of each model compared to “treat all” and “treat none” strategies.[16] Risk thresholds from 10-50% were deemed clinically relevant to inform treatment decisions, ranging from additional intravesical therapy (least aggressive) to early cystectomy (most aggressive). Bias assessments were performed to identify potential differences in model performance across clinically relevant patient-specific and disease-specific subgroups.[17]

Subset analyses

Subset analysis was performed on patients with WHO 1973 grade also available. PROGRxN-BCa was compared to the EAU risk calculator (using WHO 1973 grade),[8] and an AI model by Jobczyk et al. – the highest quality AI model identified in our previous systematic review.[18]

PROGRxN-BCa was also used to substratify intermediate-risk patients. Current guidelines recommend further categorizing these patients based on the presence of additional risk factors, including multiple tumours, tumour size > 3 cm, early recurrence (< 1 year), frequent recurrence (> 1/year), and failure of prior intravesical therapy.[12] Cumulative incidence curves for progression were compared between these guideline-defined subgroups (0, 1-2, or ≥ 3 risk factors) and thresholds derived from the top and bottom thirds of PROGRxN-BCa risk scores from the training cohort. PROGRxN-BCa was also used to substratify high-risk NMIBC into two subgroups based on the threshold corresponding to the top third of risk scores in the training cohort.

RESULTS

Overall, 1405 out of 12659 patients (11%) developed progression during a median follow-up of 3.3 years (IQR 1.6-5.8). The training cohort comprised of 3324 patients, including 1700 and 1624 patients treated at academic and community-based hospitals, respectively (**Figure 1**). During a median follow-up of 4.8 years (IQR 2.5-8.0), 459 patients (14%) from the training cohort developed progression. The international external testing cohort included 9335 patients, of which 946 patients (10%) progressed during a median follow-up of 2.8 years (IQR 1.4-5.1). Overall, the estimated five-year progression rate was 2% (IQR 1-2) for low-risk, 7% (IQR 6-8) for intermediate-risk, 21% (IQR (IQR 20-22) for high-risk, and 34% (IQR 30-39) for very high-risk patients. Patients who received guideline-concordant care had lower five-year progression rates compared to those who did not (high-risk: 18 vs 26%; very high-risk: 29 vs 42%).

The final PROGRxN-BCa model comprised of 14 features including age, sex, recurrent tumour, stage, T1 substratification, concomitant CIS, grade, variant histology, lymphovascular invasion, number of tumours, tumour diameter, repeat TURBT, BCG, and single instillation chemotherapy (SIC). Stage and grade were the most important features for predicting progression.

Overall, PROGRxN-BCa significantly outperformed the EAU risk calculator. In the training cohort, PROGRxN-BCa achieved a c-index of 0.83 (95% CI 0.81-0.84) compared to 0.76 (95% CI 0.74-0.78, p<0.001) for the EAU risk calculator. Similarly, in the external testing cohort, PROGRxN-BCa achieved a c-index of 0.79 (95% CI 0.77-0.80) compared to 0.71 (95% CI 0.70-0.72, p<0.001) for the EAU risk calculator. On bias assessment, PROGRxN-BCa outperformed the EAU risk calculator across all clinically relevant subgroups examined (**Figure 2A**). As shown in **Figure 2B**, both models were well calibrated for risks between 0-20%, however they tended to overestimate progression risk in patients with higher predicted risks. On decision curve analysis, PROGRxN-BCa had higher net benefit overall and across most subgroups for clinically relevant thresholds between 0-50% (**Figure 2C**, subgroups not shown). On subset analysis of patients with WHO 1973 grade available (n=6837), PROGRxN-BCa achieved a c-index of 0.80 (95% CI 0.78-0.81), outperforming the EAU risk calculator using the WHO 1973 grading scheme (c-index 0.74, 95% CI 0.72-0.75) and the Jobczyk AI model (c-index 0.64, 95% CI 0.63-0.66).

To understand how PROGRxN-BCa might benefit current clinical practice, we assessed how it would substratify intermediate-risk patients (n=3137) compared to current guideline recommendations. As shown in **Figure 3A**, current recommended substrata could not distinguish between the 0 and 1-2 risk factors groups, with five-year progression risks of 7 (95% CI 5-9) and 6% (95% CI 5-8), respectively. In contrast, PROGRxN-BCa separated these patients into distinct risk tertiles (**Figure 3B**). Five-year progression risks for the lower, middle, and upper tertiles were 2 (95% CI 1-4), 7 (95% CI 6-8), and 17% (95% CI 12-24), respectively, with 9% of intermediate-risk patients being reclassified into the upper tertile. PROGRxN-BCa was also capable of substratifying high-risk patients (n=5833) into two subgroups, with five-year progression risks of 11% (95% CI 9-13) and 26% (95% CI 25-28), respectively (**Figure 3C**).

CONCLUSION

In conclusion, this study demonstrates that our AI-based PROGRxN-BCa outperforms current guideline-endorsed prediction tools for NMIBC progression and improves substratification for both intermediate- and high-risk groups. In contrast to prior risk calculators, which estimate progression risk based on predefined risk groups, PROGRxN-BCa offers an individualized approach to prognostication to tailor treatment decisions for both BCG-treated and untreated patients. It also addresses limitations of prior tools by incorporating the current WHO 2004/2022 grading system and including patients who received both guideline- and non-guideline-concordant care, all while demonstrating robust performance on the largest, international NMIBC validation cohort to-date. Implementation of PROGRxN-BCa (available at https://progrxn.ca/) into NMIBC guidelines has the potential to enhance risk stratification and optimize patient management.

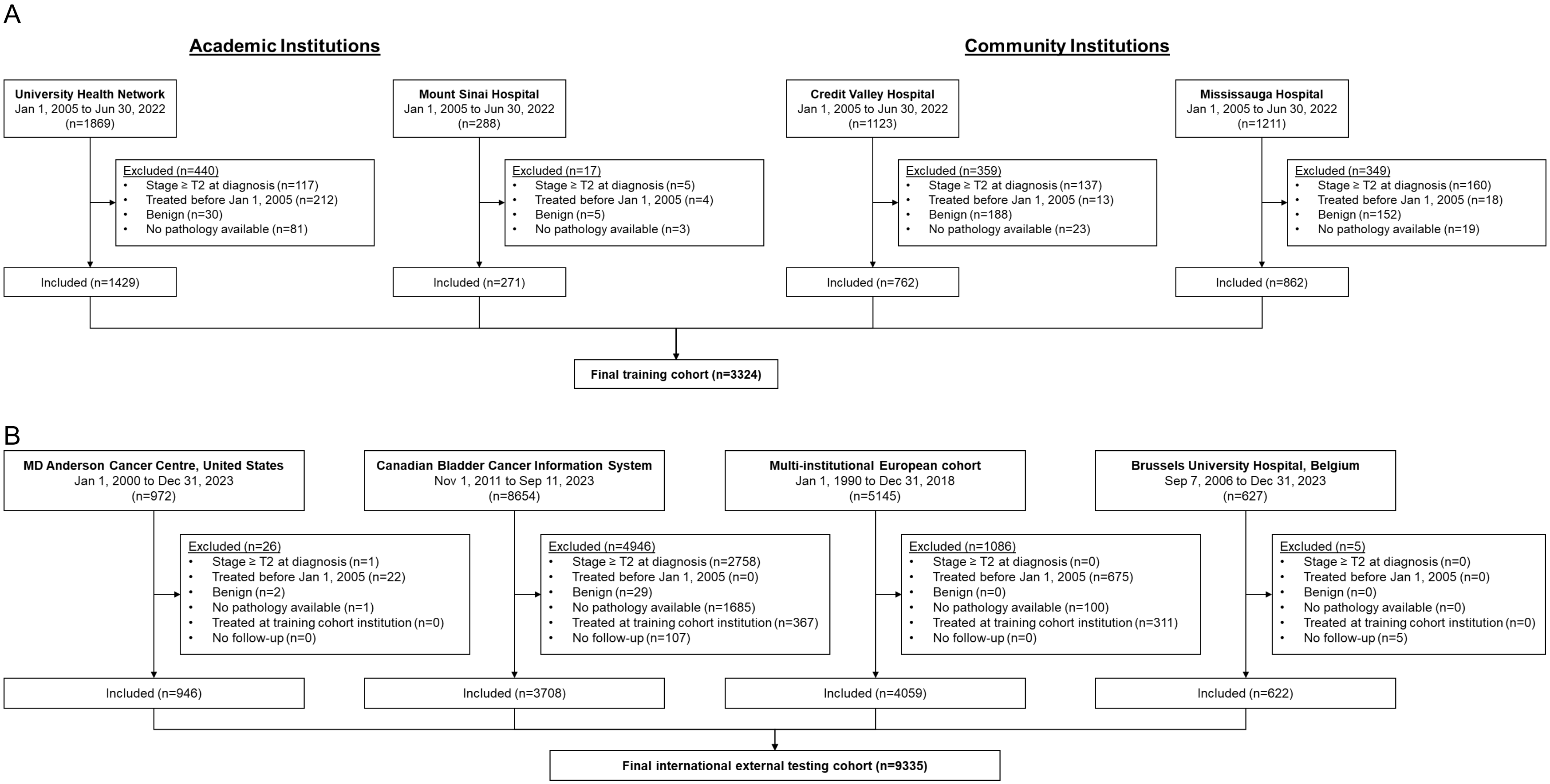
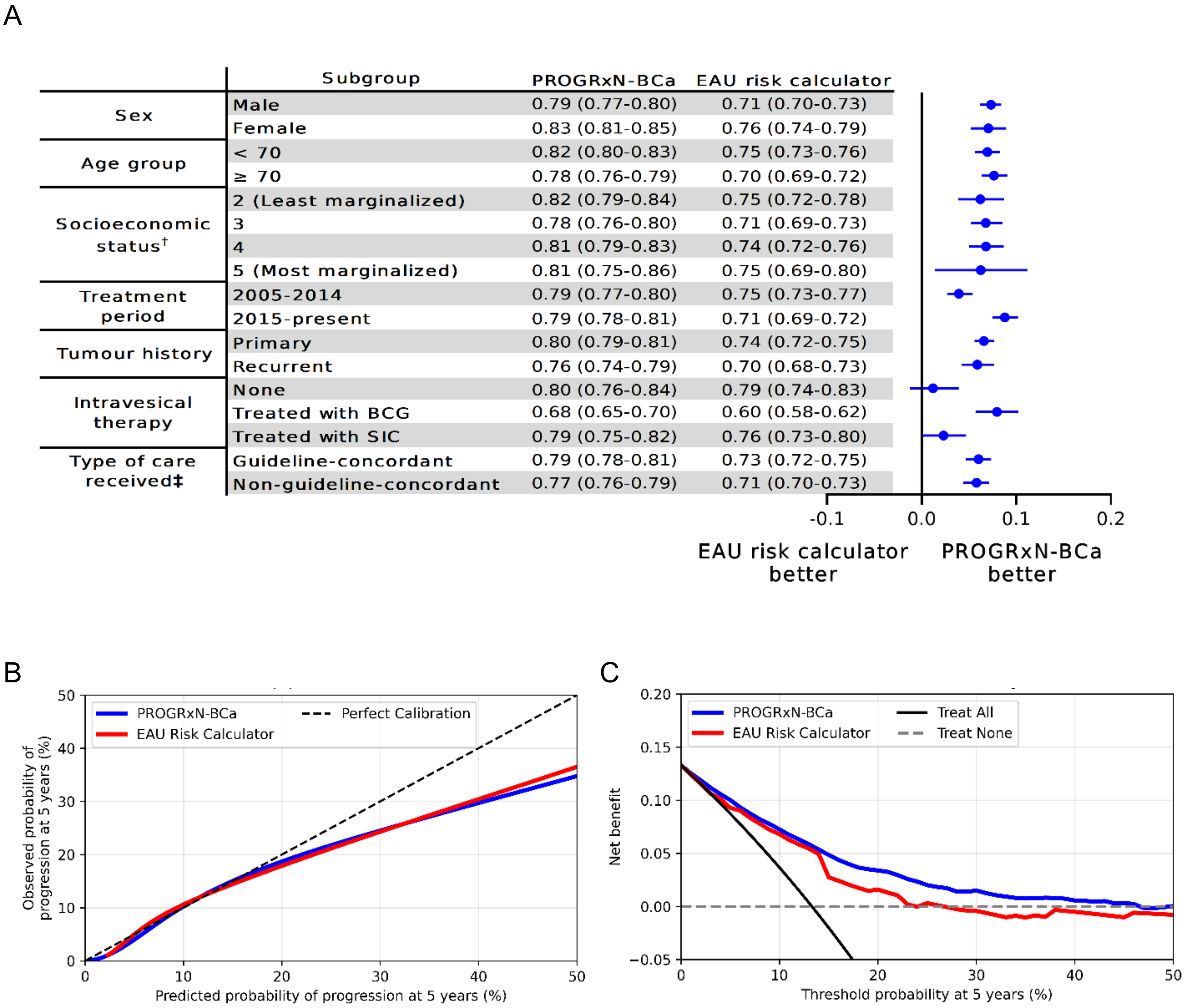
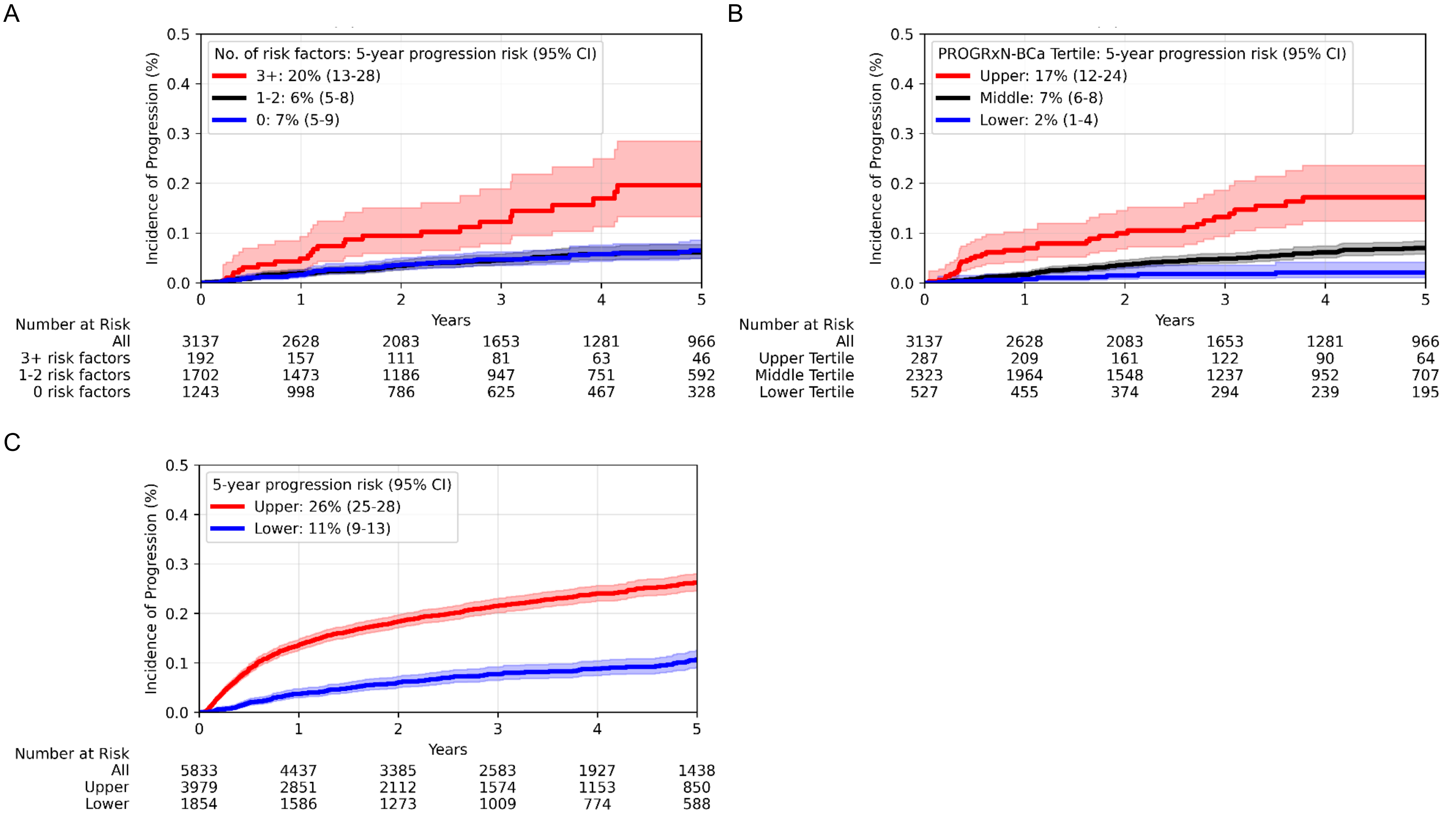

A
Academic Institutions
Community Institutions
University Health Network
Jan 1, 2005 to Jun 30, 2022
(n=1869)
Excluded (n=440)
• Stage ≥ T2 at diagnosis (n=117)
• Treated before Jan 1, 2005 (n=212)
• Benign (n=30)
• No pathology available (n=81)
Included (n=1429)
Mount Sinai Hospital
Jan 1, 2005 to Jun 30, 2022
(n=288)
Excluded (n=17)
• Stage ≥ T2 at diagnosis (n=5)
• Treated before Jan 1, 2005 (n=4)
• Benign (n=5)
• No pathology available (n=3)
Included (n=271)
Credit Valley Hospital
Jan 1, 2005 to Jun 30, 2022
(n=1123)
Excluded (n=359)
• Stage ≥ T2 at diagnosis (n=137)
• Treated before Jan 1, 2005 (n=13)
• Benign (n=188)
• No pathology available (n=23)
Included (n=762)
Mississauga Hospital
Jan 1, 2005 to Jun 30, 2022
(n=1211)
Excluded (n=349)
• Stage ≥ T2 at diagnosis (n=160)
• Treated before Jan 1, 2005 (n=18)
• Benign (n=152)
• No pathology available (n=19)
Included (n=862)
Final training cohort (n=3324)
B
MD Anderson Cancer Centre, United States
Jan 1, 2000 to Dec 31, 2023
(n=972)
Excluded (n=26)
• Stage ≥ T2 at diagnosis (n=1)
• Treated before Jan 1, 2005 (n=22)
• Benign (n=2)
• No pathology available (n=1)
• Treated at training cohort institution (n=0)
• No follow-up (n=0)
Included (n=946)
Canadian Bladder Cancer Information System
Nov 1, 2011 to Sep 11, 2023
(n=8654)
Excluded (n=4946)
• Stage ≥ T2 at diagnosis (n=2758)
• Treated before Jan 1, 2005 (n=0)
• Benign (n=29)
• No pathology available (n=1685)
• Treated at training cohort institution (n=367)
• No follow-up (n=107)
Included (n=3708)
Multi-institutional European cohort
Jan 1, 1990 to Dec 31, 2018
(n=5145)
Excluded (n=1086)
• Stage ≥ T2 at diagnosis (n=0)
• Treated before Jan 1, 2005 (n=675)
• Benign (n=0)
• No pathology available (n=100)
• Treated at training cohort institution (n=311)
• No follow-up (n=0)
Included (n=4059)
Brussels University Hospital, Belgium
Sep 7, 2006 to Dec 31, 2023
(n=627)
Excluded (n=5)
• Stage ≥ T2 at diagnosis (n=0)
• Treated before Jan 1, 2005 (n=0)
• Benign (n=0)
• No pathology available (n=0)
• Treated at training cohort institution (n=0)
• No follow-up (n=5)
Included (n=622)
Final international external testing cohort (n=9335)

Figure 1. Patient inclusion flowchart for the (A) training and (B) international external testing cohort.[19,20] The training cohort included the four Canadian academic (Princess Margaret Cancer Centre, University Health Network and Mount Sinai Hospital, Sinai Health System) and community hospitals (Credit Valley Hospital and Mississauga Hospital). The external testing cohort included MD Anderson Cancer Centre, United States; 13 academic institutions affiliated with the Canadian Bladder Cancer Information System, a prospectively maintained national database for bladder cancer patients; 15 academic institutions affiliated with the EAU NMIBC Guidelines Panel[19]; and Brussels University Hospital, Belgium. Patients that were treated at any of the training cohort institutions were excluded from the external testing cohort to prevent data leakage.



	Subgroup	PROGRxN-BCa	EAU risk calculator
Sex	Male	0.79 (0.77-0.80)	0.71 (0.70-0.73)
	Female	0.83 (0.81-0.85)	0.76 (0.74-0.79)
Age group	< 70	0.82 (0.80-0.83)	0.75 (0.73-0.76)
	≥ 70	0.78 (0.76-0.79)	0.70 (0.69-0.72)
Socioeconomic status†	2 (Least marginalized)	0.82 (0.79-0.84)	0.75 (0.72-0.78)
	3	0.78 (0.76-0.80)	0.71 (0.69-0.73)
	4	0.81 (0.79-0.83)	0.74 (0.72-0.76)
	5 (Most marginalized)	0.81 (0.75-0.86)	0.75 (0.69-0.80)
Treatment period	2005-2014	0.79 (0.77-0.80)	0.75 (0.73-0.77)
	2015-present	0.79 (0.78-0.81)	0.71 (0.69-0.72)
Tumour history	Primary	0.80 (0.79-0.81)	0.74 (0.72-0.75)
	Recurrent	0.76 (0.74-0.79)	0.70 (0.68-0.73)
Intravesical therapy	None	0.80 (0.76-0.84)	0.79 (0.74-0.83)
	Treated with BCG	0.68 (0.65-0.70)	0.60 (0.58-0.62)
	Treated with SIC	0.79 (0.75-0.82)	0.76 (0.73-0.80)
Type of care received‡	Guideline-concordant	0.79 (0.78-0.81)	0.73 (0.72-0.75)
	Non-guideline-concordant	0.77 (0.76-0.79)	0.71 (0.70-0.73)

Figure 2. (A) Bias assessment comparing subgroup-specific c-indexes for all models. Differences in c-index between PROGRxN-BCa and the EAU risk calculator are shown in the forest plot. †Socioeconomic status was only available for Canadian patients. Quintile 1 was excluded from this analysis due to insufficient samples. ‡Guideline-concordant care was defined as follows: BCG administration for high grade, T1, or any CIS; repeat TURBT for T1 tumours; and omission of BCG and repeat TURBT for primary low-risk tumours. (B) Smoothed calibration plot of all models by measuring the degree of agreement between the predicted and observed risks of progression at five years. The dotted line represents perfect calibration. (C) Clinical utility of all models assessed using decision curve analysis. The higher the net benefit, the better the potential clinical impact and identification of patients who may benefit most from treatment intensification.



Number at Risk	0	1	2	3	4	5
All	3137	2628	2083	1653	1281	966
3+ risk factors	192	157	111	81	63	46
1-2 risk factors	1702	1473	1186	947	751	592
0 risk factors	1243	998	786	625	467	328

Number at Risk	0	1	2	3	4	5
All	3137	2628	2083	1653	1281	966
Upper Tertile	287	209	161	122	90	64
Middle Tertile	2323	1964	1548	1237	952	707
Lower Tertile	527	455	374	294	239	195

Number at Risk	0	1	2	3	4	5
All	5833	4437	3385	2583	1927	1438
Upper	3979	2851	2112	1574	1153	850
Lower	1854	1586	1273	1009	774	588

Figure 3. Substratification of intermediate-risk patients (n=3137) based on (A) the number of risk factors (0, 1-2, or ≥ 3 risk factors) and (B) PROGRxN-BCa tertiles. Risk factors included multiple tumours, tumour size > 3 cm, early recurrence (< 1 year), frequent recurrence (> 1/year), and failure of prior intravesical therapy. PROGRxN-BCa separated patients into risk tertiles based on cutoffs at the top and bottom thirds of risk scores from the training cohort. (C) Substratification of high-risk NMIBC (n=5833). Patients were substratified into "Upper" and "Lower" subgroups based on the threshold corresponding to the top third of PROGRxN-BCa risk scores from the training cohort. This threshold is the same one used to define the "Upper Risk Tertile" for intermediate-risk NMIBC. The estimated five-year progression risk (95% confidence interval) for each subgroup is indicated.

REFERENCES

1. Kwong JCC, Wu J, Malik S, Khondker A, Gupta N, Bodnariuc N, et al. Predicting non-muscle invasive bladder cancer outcomes using artificial intelligence: a systematic review using APPRAISE-AI. NPJ Digit Med. 2024 Apr 18;7(1):98.
2. Kamat AM, Hahn NM, Efstathiou JA, Lerner SP, Malmström PU, Choi W, et al. Bladder cancer. The Lancet. 2016 Dec 3;388(10061):2796–810.
3. Compérat E, Amin MB, Cathomas R, Choudhury A, De Santis M, Kamat A, et al. Current best practice for bladder cancer: a narrative review of diagnostics and treatments. The Lancet. 2022 Nov 12;400(10364):1712–21.
4. Necchi A, Roumiguié M, Kamat AM, Shore ND, Boormans JL, Esen AA, et al. Pembrolizumab monotherapy for high-risk non-muscle-invasive bladder cancer without carcinoma in situ and unresponsive to BCG (KEYNOTE-057): a single-arm, multicentre, phase 2 trial. The Lancet Oncology. 2024 Jun 1;25(6):720–30.
5. Fernandez-Gomez J, Madero R, Solsona E, Unda M, Martinez-Piñeiro L, Gonzalez M, et al. Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the CUETO scoring model. J Urol. 2009 Nov;182(5):2195–203.
6. Sylvester RJ, van der Meijden APM, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. Eur Urol. 2006 Mar;49(3):466–465; discussion 475-477.
7. Villegas E, Lajkosz K, Din S, Kuk C, Chan A, Kwong JCC, et al. Long-Term Recurrence Risk, Metastatic Potential, and Length of Cystoscopic Surveillance of Low-Grade Non-Muscle-Invasive Bladder Cancer. J Urol. 2024 Oct 29;101097JU0000000000004305.
8. Sylvester RJ, Rodríguez O, Hernández V, Turturica D, Bauerová L, Bruins HM, et al. European Association of Urology (EAU) Prognostic Factor Risk Groups for Non–muscle-invasive Bladder Cancer (NMIBC) Incorporating the WHO 2004/2016 and WHO 1973 Classification Systems for Grade: An Update from the EAU NMIBC Guidelines Panel. European Urology. 2021 Apr 1;79(4):480–8.
9. Lobo N, Hensley PJ, Bree KK, Nogueras-Gonzalez GM, Navai N, Dinney CP, et al. Updated European Association of Urology (EAU) Prognostic Factor Risk Groups Overestimate the Risk of Progression in Patients with Non-muscle-invasive Bladder Cancer Treated with Bacillus Calmette-Guérin. Eur Urol Oncol. 2022 Feb;5(1):84–91.
10. Kwong JCC, Khondker A, Meng E, Taylor N, Kuk C, Perlis N, et al. Development, multi-institutional external validation, and algorithmic audit of an artificial intelligence-based Side-specific Extra-Prostatic Extension Risk Assessment tool (SEPERA) for patients undergoing radical prostatectomy: a retrospective cohort study. The Lancet Digital Health. 2023 Jul 1;5(7):e435–45.
11. Kwong JCC, Khondker A, Lajkosz K, McDermott MBA, Frigola XB, McCradden MD, et al. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. JAMA Network Open. 2023 Sep 25;6(9):e2335377.
12. Tan WS, Steinberg G, Witjes JA, Li R, Shariat SF, Roupret M, et al. Intermediate-risk Non-muscle-invasive Bladder Cancer: Updated Consensus Definition and Management Recommendations from the International Bladder Cancer Group. Eur Urol Oncol. 2022 Oct;5(5):505–16.
13. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2019 Mar 30;38(7):1276–96.

14. Kwong JCC, McLoughlin LC, Haider M, Goldenberg MG, Erdman L, Rickard M, et al. Standardized Reporting of Machine Learning Applications in Urology: The STREAM-URO Framework. Eur Urol Focus. 2021 Jul;7(4):672–82.
15. Jarrett D, Cebere BC, Liu T, Curth A, Schaar M van der. HyperImpute: Generalized Iterative Imputation with Automatic Model Selection. In: Proceedings of the 39th International Conference on Machine Learning [Internet]. PMLR; 2022 [cited 2023 Nov 23]. p. 9916–37. Available from: https://proceedings.mlr.press/v162/jarrett22a.html
16. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
17. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health. 2022 May;4(5):e384–97.
18. Jobczyk M, Stawiski K, Kaszkowiak M, Rajwa P, Różański W, Soria F, et al. Deep Learning-based Recalibration of the CUETO and EORTC Prediction Tools for Recurrence and Progression of Non-muscle-invasive Bladder Cancer. Eur Urol Oncol. 2022 Feb;5(1):109–12.
19. van Rhijn BWG, Hentschel AE, Bründl J, Compérat EM, Hernández V, Čapoun O, et al. Prognostic Value of the WHO1973 and WHO2004/2016 Classification Systems for Grade in Primary Ta/T1 Non-muscle-invasive Bladder Cancer: A Multicenter European Association of Urology Non-muscle-invasive Bladder Cancer Guidelines Panel Study. Eur Urol Oncol. 2021 Apr;4(2):182–91.
20. Black PC, Alimohamed N, Kassouf W, Gore JL, McCoy KD, Nelson BH, et al. Building the Canadian Bladder Cancer Research Network (CBCRN): Progress during a pandemic. Can Urol Assoc J. 2022 Jun;16(6):E307–14.

ASSOCIATIONS BETWEEN HIGH-SENSITIVITY CARDIAC TROPONIN I AND 30-DAY MORTALITY FOLLOWING OFF-PUMP AND ON-PUMP CORONARY ARTERY BYPASS SURGERY: A SUBANALYSIS OF THE VISION CARDIAC SURGERY STUDY

Grace S Lee[1], Derrick Y Tam[2], Dominique Vervoort[3], Shun Fu Lee[4], Katheryn Brady[4], Richard Whitlock[4,5], Emilie Belley-Cote[5], Andre Lamy[5], PJ Devereaux[4,5], Stephen E Fremes[2,3]

[1]Division of Cardiac Surgery, University of Toronto, Toronto, Ontario, Canada
[2]Schulich Heart Program, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Ontario, Canada
[3]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
[4]Population Health Research Institute, Hamilton, Ontario, Canada
[5]McMaster University, Hamilton, Ontario, Canada

INTRODUCTION

Coronary artery bypass surgery (CABG) remains the gold standard treatment for extensive coronary disease. Compared to medical therapy and percutaneous coronary intervention (PCI), CABG provides significant long-term mortality benefits, symptom reduction and protection against non-fatal cardiac events.[1] However, it is also an invasive procedure that often involves manipulating the heart and disruption of physiological coronary blood flow. A transient rise in cardiac enzymes, such as creatine-kinase myocardial band and troponin I and T, is routinely observed in patients after CABG. This may result from cellular pathways separate from myocardial necrosis and typically poses no major effect on long-term outcomes.[2] However, extremely high levels of ischemic biomarkers are concerning for irreversible myocardial injury from complications such as graft failure, poor myocardial protection, incomplete revascularization, or tissue damage related to the surgery itself.[3]

High-sensitivity cardiac troponin I (hs-cTnI) is one biomarker of choice used to diagnose spontaneous myocardial infarction (MI).[4] Although small elevations in hs-cTnI are prognostically important after non-cardiac surgery, high hs-cTnI values in excess of usual diagnostic thresholds for MI are routinely observed after cardiac surgery.[5–7] There is currently no clear consensus on the clinically significant level of hs-cTnI after CABG that should indicate additional workup or intervention.[2] The hs-cTnI threshold defining postoperative MI has also provoked widespread controversy for its impact on outcomes reporting of major clinical trials comparing PCI to CABG, given the starkly different levels of hs-cTnI observed after these procedures.[8]

The Vascular Events in Surgery Patients Cohort Evaluation (VISION) Cardiac Surgery Study examined routine hs-cTnI levels associated with clinically important myocardial injury after cardiac surgery. The hs-cTnI threshold associated with significant adverse outcomes varied according to the cardiac surgical procedure performed; however, the VISION study combined CABG and aortic valve replacement surgeries in their original analysis and did not report a threshold value for isolated CABG.[9] Furthermore, it did not differentiate between CABG performed with cardiopulmonary bypass (ONCAB) and without cardiopulmonary bypass (OFFCAB).[9] OFFCAB is a fundamentally distinct procedure with typically lower postoperative hs-cTnI levels. Yet, numerous studies have found comparable, or even inferior, long-term outcomes in OFFCAB compared to ONCAB.[10–12] It remains unclear what the clinically relevant hs-cTnI threshold should be in patients undergoing isolated CABG and whether the same threshold should be used for OFFCAB and ONCAB.

Objective

We aimed to determine hs-cTnI thresholds associated with increased 30-day mortality and major vascular complications (MVCs) in patients undergoing isolated CABG and whether these values differed between those undergoing OFFCAB and ONCAB.

METHODS

Study Design

We conducted a secondary analysis of patients who underwent isolated CABG in the VISION Cardiac Surgery study, a multicentre international prospective cohort study conducted from 2013-2019. Patients were excluded from this analysis if they had: preoperative MI within 24 hours; preoperative hs-cTnI ≥300 ng/L within 12 hours; salvage cardiac surgery; or any other concomitant cardiac surgery. Baseline variables included demographics, cardiovascular risk factors, EuroSCORE II, New York Heart Association (NYHA) class and Canadian Cardiovascular Society (CCS) class of angina. Hs-cTnI levels on postoperative day 1 (POD1) were measured using a standardized assay with an upper reference limit (URL) of 26 ng/L and compared between patients who underwent OFFCAB and ONCAB.

Outcomes

The primary outcome of interest was 30-day all-cause mortality. The secondary outcome of interest was 30-day MVCs, a composite of vascular mortality, postoperative MI and insertion of a mechanical assist device.

Statistical Analysis

All statistical analyses were conducted in R 4.1.0 (R Foundation for Statistical Computing, Vienna). Baseline characteristics and 30-day outcomes were compared between OFFCAB and ONCAB using Student's t-test or Kruskal-Wallis test. Box and whisker plots were constructed to compare hs-cTnI levels according to operation and outcomes. Patients without recorded POD1 hs-cTnI or hs-cTnI <100 ng/L or >51,000 ng/L were excluded from Cox regression analyses.

Cox regression analysis was used to determine the hazard ratios (HRs) for 30-day all-cause mortality and MVCs as a function of log-transformed POD1 peak hs-cTnI. Models were adjusted by EuroSCORE II (predicted risk of 30-day mortality), with OFFCAB versus ONCAB as an interaction term. Fitted models that predicted HRs as a function of log-hs-cTnI were derived via natural cubic spline regression. From these fitted models, the lowest hs-cTnI threshold associated with HR≥1.00 was identified for all CABG patients and compared between OFFCAB and ONCAB patients. The 95% confidence intervals (95%CI) surrounding these hs-cTnI thresholds were determined where possible via bootstrap sampling.

RESULTS

Baseline Characteristics

The VISION Cardiac Surgery Study included 13,862 patients, of which 6,505 (OFFCAB=1,141, ONCAB=5,364) underwent isolated CABG and were eligible for this subanalysis. The median age was 65.5 (IQR 58.6-71.8) years, and 80.7% of patients were men. Patients undergoing OFFCAB had a lower median EuroSCORE II (1.1% [IQR 0.7-1.8]) than ONCAB (1.2% [IQR 0.8-1.9]) (p=0.002) and were less likely to have had a previous MI (40.3% vs 51.3%, p<0.001). Patients undergoing OFFCAB presented with less advanced NYHA status (p<0.001) and CCS angina class (p<0.001) compared to ONCAB **(Table 1)**.

OFFCAB versus ONCAB: 30-Day Outcomes

The median POD1 peak hs-cTnI level after isolated CABG was recorded across 6,382 patients to be 2,446 ng/L (IQR 1,164-5,654). The peak hs-cTnI was significantly lower after OFFCAB (640 ng/L [IQR 264-1,689]) than ONCAB (2,972 ng/L [IQR 1,536-6,448], p<0.001). Incomplete revascularization rates were higher in OFFCAB (26.3%) than ONCAB (11.2%, p<0.001). There was no difference in 30-day all-cause mortality between OFFCAB and ONCAB (1.7% vs 1.4%, p=0.5). Similarly, there was no difference in MVCs between OFFCAB (2.1%) and ONCAB (2.3%) (p=0.7), nor its individual components of vascular death, MI or insertion of a mechanical assist device (**Table 2).**

Log-Peak hs-cTnI Thresholds and 30-Day Mortality

Increased log-peak hs-cTnI was associated with a higher likelihood of 30-day mortality for all isolated CABG patients (HR=1.7 [95%CI 1.4-2.1) after adjusting for EuroSCORE II. OFFCAB had a significant positive interaction effect on the association between hs-cTnI and 30-day mortality (p=0.002). Amongst all patients who underwent isolated CABG, the lowest hs-cTnI threshold associated with a HR≥1.00 for 30-day mortality was 6,549 ng/L (95%CI 3,609-8,381). The hs-cTnI threshold associated with increased risk of 30-day mortality after OFFCAB was ≥4,708 ng/L (95%CI 581-7,177), compared to ≥6,806 ng/L (95%CI 4,001-13,993) after ONCAB **(Figure 1A-C).**

Log-Peak hs-cTnI Thresholds and 30-Day Major Vascular Complications

Similarly, increased log-peak hs-cTnI was associated with an overall higher likelihood of MVCs amongst all patients (adjusted HR=1.7 [95%CI 1.5-1.9]). OFFCAB had a significant positive interaction effect on this association (p=0.01). The hs-cTnI threshold associated with HR≥1.00 for MVC after isolated CABG was 4,781 ng/L (95%CI 1,470-7,557). This hs-cTnI threshold was lower for patients who underwent OFFCAB (4,062 ng/L) compared to ONCAB (6,912 ng/L) **(Figure 1D-F)**.

CONCLUSIONS

We conducted a secondary analysis of the landmark VISION Cardiac Surgery study comparing hs-cTnI levels in patients after isolated OFFCAB and ONCAB. Our findings (i) provide an evidence-based and nuanced definition of clinically significant hs-cTnI elevation to guide medical/surgical intervention after CABG and (ii) shed light on the physiology behind troponin elevations associated with cardiopulmonary bypass.

One of the most worrisome complications of CABG is postoperative MI, which could be caused by graft failure, poor myocardial protection or direct surgical trauma.[13] Existing evidence shows that, even in the absence of supportive ECG or imaging findings, hs-cTnI is a strong prognosticator of adverse outcomes.[5–7] However, routine collection of hs-cTnI after CABG varies amongst hospitals as the threshold for postoperative MI remains contested.[2] Guideline definitions range from >10x URL to >70x URL, without clear evidence behind these thresholds or consideration that postoperative hs-cTnI levels are procedure-dependent.[2,14] These definitions have also been applied to measure outcomes in major clinical trials to much debate. The Evaluation of XIENCE versus CABG for Effectiveness of Left Main Revascularization (EXCEL) trial compared PCI to CABG using a biomarker-only threshold of hs-cTnI >10x URL to define perioperative MI in both groups.[15] This generated significant controversy regarding the overdiagnosis of perioperative MI in the CABG group and biased outcomes that favoured PCI for treatment of left main disease.[8,16] Our VISION subanalysis provides an evidence-based definition of clinically significant myocardial injury after isolated CABG that far exceeds current thresholds at 6,549 ng/L (252x URL) for all-cause mortality and 4,781 ng/L (189x URL) for MVCs; these thresholds also far exceed those used to define MI in the EXCEL trial.

Establishing nuanced definitions for clinically significant troponin leak between OFFCAB and ONCAB also helps to understand the physiology of cardiopulmonary bypass. While the hs-cTnI thresholds for mortality HR≥1.00 exceeded those defined by current guidelines, ONCAB had a significantly higher hs-cTnI threshold (6,806 ng/L, 261x URL) than OFFCAB (4,708 ng/L, 181x URL), with similar findings for MVCs. This may indicate that the excess hs-cTnI typically associated with cardiopulmonary bypass arises from reversible and nonfatal forms of myocardial injury, such as cell edema, cannulation incisions, and reperfusion-induced free radical release.[13]

In summary, we conducted a subanalysis of isolated CABG procedures in the landmark VISION Cardiac Surgery study. We found that the hs-cTnI thresholds associated with increased 30-day mortality and MVCs are substantially lower in patients undergoing OFFCAB compared to ONCAB. Our findings represent a validated and procedure-specific definition of clinically significant hs-cTnI thresholds that can be used routinely to prognosticate the need for early reintervention and clinical decision-making after OFFCAB and ONCAB.

Table 1. Baseline characteristics of all CABG and ONCAB versus OFFCAB cohorts

	Overall (n=6505)	**ONCAB (n=5364)**	**OFFCAB (n=1141)**	**p-value**
Age (median [IQR])	65.5 [58.6, 71.8]	65.5 [58.7, 71.8]	65.8 [58.2, 72.0]	0.7
Female sex (%)	1258 (19.3)	1061 (19.8)	197 (17.3)	0.06
Baseline BMI (median [IQR])	28.0 [25.2, 31.2]	28.1 [25.2, 31.4]	27.3 [24.8, 30.7]	**<0.001**
Hypertension (%)	5090 (78.2)	4234 (78.9)	856 (75.0)	**0.004**
Diabetes (%)	2633 (40.5)	2220 (41.4)	413 (36.2)	**0.001**
COPD (%)	513 (7.9)	426 (7.9)	87 (7.6)	0.8
Tobacco use (%)	3870 (59.5)	3224 (60.1)	646 (56.6)	**0.03**
Previous MI (%)	3211 (49.4)	2751 (51.3)	460 (40.3)	**<0.001**
MI within 90 days (%)	1769 (27.2)	1562 (29.1)	207 (18.1)	**<0.001**
Heart failure (%)	794 (12.2)	673 (12.5)	121 (10.6)	0.08
Previous stroke (%)	339 (5.2)	289 (5.4)	50 (4.4)	0.2
Atrial Fibrillation (%)	438 (6.7)	376 (7.0)	62 (5.4)	0.07
Peripheral arterial disease (%)	600 (9.2)	477 (8.9)	123 (10.8)	0.05
CCS class (%)				**<0.001**
0	789 (12.1)	620 (11.6)	169 (14.8)	
1	890 (13.7)	655 (12.2)	235 (20.6)	
2	2018 (31.0)	1662 (31.0)	356 (31.2)	
3	1969 (30.3)	1702 (31.7)	267 (23.4)	
4	839 (12.9)	725 (13.5)	114 (10.0)	
NYHA class (%)				**<0.001**
0	1108 (17.0)	819 (15.3)	289 (25.3)	
1	1210 (18.6)	1002 (18.7)	208 (18.2)	
2	2246 (34.5)	1841 (34.3)	405 (35.5)	
3	1497 (23.0)	1292 (24.1)	205 (18.0)	
4	444 (6.8)	410 (7.6)	34 (3.0)	
Urgency (%)				ns
Urgent	1881 (28.9)	1629 (30.4)	252 (22.1)	
Emergent	97 (1.5)	81 (1.5)	16 (1.4)	
Elective	4527 (69.6)	3654 (68.1)	873 (76.5)	
Poor mobility (%)	248 (3.8)	200 (3.7)	48 (4.2)	0.5
Baseline Creatinine (µmol/L) (mean [SD])	84.2 (78.8)	85.9 (83.7)	76.6 (49.6)	**<0.001**
EuroSCORE II (median [IQR])	1.2 [0.8, 1.9]	1.2 [0.8, 1.9]	1.1 [0.7, 1.8]	**0.002**

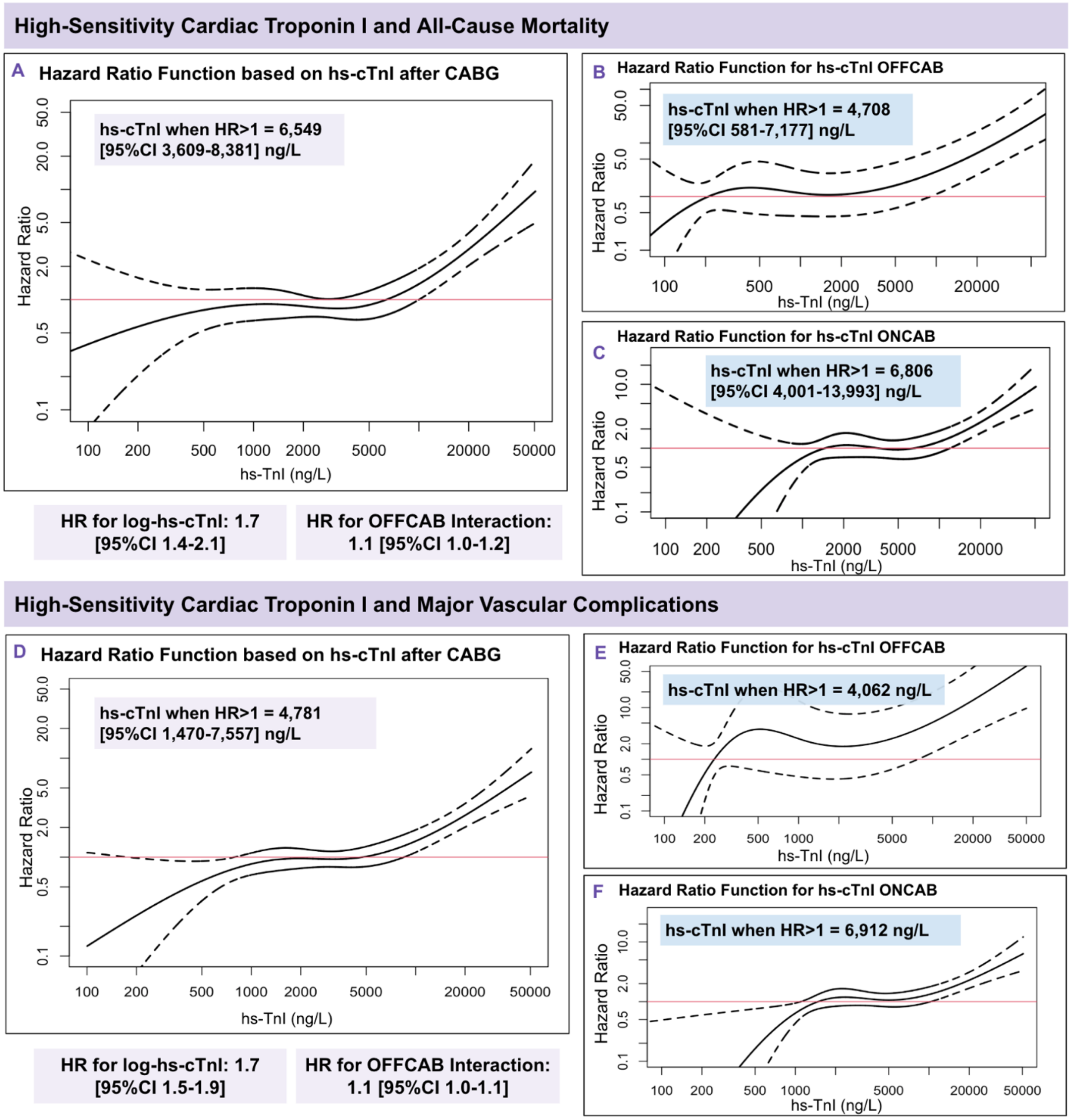
Abbreviations: BMI = body mass index, CCS = Canadian Cardiovascular Society, MI = myocardial infarction, NYHA = New York Heart Association, ns = non-significant, OFFCAB = off-pump CABG, ONCAB = on-pump CABG.

Table 2: Peak hs-cTnI levels on postoperative day 1 and 30-day outcomes after isolated CABG

	Overall (n=6505)	**ONCAB (n=5364)**	**OFFCAB (n=1141)**	**p-value**
Peak hs-cTnI Day 1 (median [IQR])	2445.8 [1164.2-5653.5]	2972 [1536.5-6448.8]	640 [264-1688.5]	**<0.001**
Complete revascularization (%)	5594 (86.1)	4755 (88.8)	839 (73.7)	**<0.001**
All-cause mortality (%)	92 (1.4)	73 (1.4)	19 (1.7)	0.5
Vascular death (%)	72 (1.1)	57 (1.1)	15 (1.3)	0.6
Mechanical assist device insertion (%)	75 (1.2)	67 (1.2)	8 (0.7)	0.2
MVC (%)	150 (2.3)	126 (2.3)	24 (2.1)	0.7
MI (%)	23 (0.4)	17 (0.3)	6 (0.5)	0.4

Abbreviations: MI = myocardial infarction, MVC = major vascular complications, OFFCAB = off-pump CABG, ONCAB = on-pump CABG.

Figure 1: Associations between hs-cTnI and 30-day **(1A-C)** all-cause mortality and **(1D-F)** major vascular complications in isolated CABG, OFFCAB and ONCAB.



Abbreviations: 95% CI = 95% confidence interval, HR = hazard ratio, OFFCAB = off-pump CABG, ONCAB = on-pump CABG.

References

1. Lawton JS, Tamis-Holland JE, Bangalore S, et al. 2021 ACC/AHA/SCAI Guideline for Coronary Artery Revascularization. *Circulation*. 2022;145:e18–e114.

2. Gaudino M, Flather M, Capodanno D, et al. European Association of Cardio-Thoracic Surgery (EACTS) expert consensus statement on perioperative myocardial infarction after cardiac surgery. *European Journal of Cardio-Thoracic Surgery*. 2024;65.

3. Gaudino M, Dangas GD, Angiolillo DJ, et al. Considerations on the Management of Acute Postoperative Ischemia After Cardiac Surgery: A Scientific Statement From the American Heart Association. *Circulation*. 2023;148:442–454.

4. Ruetzler K, Smilowitz NR, Berger JS, et al. Diagnosis and Management of Patients With Myocardial Injury After Noncardiac Surgery: A Scientific Statement From the American Heart Association. *Circulation*. 2021;144.

5. Heuts S, Denessen EJS, Daemen JHT, et al. Meta-Analysis Evaluating High-Sensitivity Cardiac Troponin T Kinetics after Coronary Artery Bypass Grafting in Relation to the Current Definitions of Myocardial Infarction. *Am J Cardiol*. 2022;163:25–31.

6. Pölzl L, Engler C, Sterzinger P, et al. Association of High-Sensitivity Cardiac Troponin T With 30-Day and 5-Year Mortality After Cardiac Surgery. *J Am Coll Cardiol*. 2023;82:1301–1312.

7. Omran H, Deutsch MA, Groezinger E, et al. High-sensitivity cardiac troponin I after coronary artery bypass grafting for post-operative decision-making. *Eur Heart J*. 2022;43:2388–2403.

8. Ruel M, Falk V, Farkouh ME, et al. Myocardial Revascularization Trials. *Circulation*. 2018;138:2943–2951.

9. Devereaux PJ, Lamy A, Chan MTV, et al. High-Sensitivity Troponin I after Cardiac Surgery and 30-Day Mortality. *New England Journal of Medicine*. 2022;386:827–836.

10. Diegeler A, Börgermann J, Kappert U, et al. Five-Year Outcome After Off-Pump or On-Pump Coronary Artery Bypass Grafting in Elderly Patients. *Circulation*. 2019;139:1865–1871.

11. Quin JA, Wagner TH, Hattler B, et al. Ten-Year Outcomes of Off-Pump vs On-Pump Coronary Artery Bypass Grafting in the Department of Veterans Affairs. *JAMA Surg*. 2022;157:303.

12. Lamy A, Devereaux PJ, Prabhakaran D, et al. Five-Year Outcomes after Off-Pump or On-Pump Coronary-Artery Bypass Grafting. *New England Journal of Medicine*. 2016;375:2359–2368.

13. Schneider U, Mukharyamov M, Beyersdorf F, et al. The value of perioperative biomarker release for the assessment of myocardial injury or infarction in cardiac surgery. *European Journal of Cardio-Thoracic Surgery*. 2022;61:735–741.

14. Thygesen K, Alpert JS, Jaffe AS, et al. Fourth Universal Definition of Myocardial Infarction (2018). *Circulation*. 2018;138.

15. Stone GW, Sabik JF, Serruys PW, et al. Everolimus-Eluting Stents or Bypass Surgery for Left Main Coronary Artery Disease. *New England Journal of Medicine*. 2016;375:2223–2235.

16. Ben-Yehuda O, Chen S, Redfors B, et al. Impact of large periprocedural myocardial infarction on mortality after percutaneous coronary intervention and coronary artery bypass grafting for left main disease: an analysis from the EXCEL trial. *Eur Heart J*. 2019;40:1930–1941.

February 26th, 2025

Dr. Michael G. Fehlings, Vice Chair Research
Dr. Andras Kapus, Associate Vice Chair Research
Renata Musa, Research Program Coordinator
Department of Surgery, University of Toronto

Dear Dr. Fehlings, Dr. Kapus and Ms. Musa,

On behalf of my coauthors, it is my pleasure to submit our abstract titled “Associations between High-Sensitivity Cardiac Troponin I and 30-day Mortality following Off-pump and On-pump Coronary Artery Bypass Surgery” to the 2025 Gallie Day Oral Presentation Competition.

High-sensitivity cardiac troponin is the gold standard for diagnosing spontaneous myocardial infarction. However, the threshold of concern for high-sensitivity cardiac troponin I (hs-cTnI) elevations after coronary artery bypass surgery (CABG) remains unknown. We have conducted a subanalysis of the landmark VISION Cardiac Surgery Study, a multicentre international cohort study of 13,682 patients, to determine the threshold of hs-cTnI associated with mortality after isolated CABG and compared this between on-pump CABG and off-pump CABG.

This study was conducted under the supervision of Dr. Stephen E. Fremes with the Division of Cardiac Surgery at the University of Toronto from September 2024 to present. It was performed in partnership with the original VISION Cardiac Surgery Study authors from the Population Health Research Institute at McMaster University, who provided the raw data and approved of this secondary analysis. The first author contributed to study development, and carried out the statistical analyses, interpretation of results and manuscript preparation. The second, third and senior authors directly supervised and contributed to the study design, data analysis and manuscript preparation.

We thank you for your consideration and the opportunity to participate in the 2025 Gallie Day Oral Presentation Competition.

Warm regards,

Grace S Lee, MD

Resident
Division of Cardiac Surgery
Department of Surgery
University of Toronto

LATENT HOME TIME TRAJECTORIES FOR ISOLATED MODERATE TO SEVERE TRAUMATIC BRAIN INJURY SURVIVORS: A DATA-DRIVEN METHOD IDENTIFIES DISTINCT OUTCOME PHENOTYPES

Armaan K Malhotra MD (SSTP)[1,2], Avery B Nathens[3], Husain Shakil[1,2],
Vishwathsen Karthikeyan[1,2], Christopher S Lozano[1,2], Jetan H Badhiwala[4],
Francois Mathieu[1,5], Benjamin Davidson[4], Yingshi He[1], Abhaya V Kulkarni[2,6],
Christopher D Witiw[1,2], Kevin E Thorpe[7], Jefferson R Wilson[1,2]
[1]Division of Neurosurgery, Unity Health Toronto, Toronto, Ontario
[2]Institute for Health Policy Management & Evaluation, University of Toronto, Toronto, Ontario
[3]Division of General Surgery, Sunnybrook Health Sciences Center, Toronto, Ontario
[4]Division of Neurosurgery, Sunnybrook Health Sciences Center, Toronto, Ontario
[5]Interdepartmental Division of Critical Care, University of Toronto, Toronto, Ontario
[6]Division of Neurosurgery, Hospital for Sick Children, Toronto, Ontario
[7]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario

The authors have decided not to make the research results available at this time and will provide updates as soon as the results can be shared.

.

WHAT IT MEANS TO HAVE A SCAR: INSIGHTS FROM 20,000 PEOPLE

Whitney Quong (SSTP)[1], Cornelia Borkhoff[2], Aaron Drucker[3], Joel Fish[1]
[1]Division of Plastic, Reconstructive, and Aesthetic Surgery, Department of Surgery, University of Toronto, Toronto, Ontario
[2]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario
[3]Division of Dermatology, Department of Medicine, University of Toronto, Toronto, Ontario

The authors have decided not to make the research results available at this time and will provide updates as soon as the results can be shared.

ENDOTHELIN-1 AS A BIOMARKER IN HUMAN EX VIVO LUNG PERFUSATE TO PREDICT DONOR LUNG SUITABILITY

Abby McCaig[1,2], Xuanzi Zhou[1,2], Thomas Borrillo[1], Marcelo Cypel[1-3], Shaf Keshavjee[1-3], Andrew Sage[1-3*], Mingyao Liu[1-3*]

[1]Latner Thoracic Research Laboratories, Toronto General Hospital Research Institute, University Health Network, Toronto, Ontario, Canada.

[2]Institute of Medical Science, Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

[3]Department of Surgery, Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

*Co-senior authors

INTRODUCTION

Lung transplantation (LTx) is the only curative treatment option for patients suffering from end-stage lung disease.[1] Although, the number of patients requiring an LTx exceeds the amount of donor lungs available.[2] This can be partially attributed to low donor lung utilization rates, driven by concerns about donor lung quality and its potential impact on recipient outcomes.[2] Development of primary graft dysfunction (PGD) within the first 72 hours following LTx is the leading cause of mortality post-LTx.[3] The ex vivo lung perfusion (EVLP) system was developed to evaluate and recondition marginal donor lungs for LTx, by mimicking physiological conditions through perfusate circulation and mechanically ventilating the lungs.[4,5] The perfusion of lungs outside the body allows for advanced assessments of marginal donor lungs to evaluate their transplant suitability.[5] Advanced assessments performed on EVLP include unobstructed X-rays,[6,7] and biochemical,[8] physiological,[9] and biological assessments.[10,11] Currently, biological assessments are not a part of standard clinical EVLP assessment, yet research has shown biomarkers in EVLP perfusate can provide information on the donor lungs inflammatory responses,[11–14] cell death processes,[15–17] and endothelial damage.[18–20] Although, current biomarker research must overcome many limitations prior to clinical use. Preliminary research on endothelin-1 (ET-1), a potent chemokine that stimulates rapid vasoconstriction, suggests that levels during EVLP may reflect donor lung endothelial damage and be used to predict donor lung outcomes.[18,21] Therefore, we examined ET-1 to understand the mechanism of donor lung injury, predict donor lung outcomes, and overcome the current limitations of biomarker research.

METHODS

Tissue sample collection & pathway enrichment analysis

Peripheral donor lung tissue samples were collected between 2011 and 2015 for n = 88 cases. The first sample was collected during donor lung preservation while on ice, referred to as pre-EVLP. The donor lungs are then placed on EVLP for 4-6 hours. The donor lungs are placed back on ice following EVLP, and a second biopsy is taken, which is referred to as post-EVLP. All samples were snap-frozen in liquid nitrogen. RNA was isolated from tissue samples using the RNeasy Mini Kit (Qiagen Canada, Toronto, Canada). A nanodrop spectrophotometer (Thermo Fisher Scientific Canada, Ottawa, Canada) was used to assess RNA quality. The samples with an RNA Integrity Number above 7.0 and a concentration above 100 ng/ul proceeded to the next steps. Following the manufacturer's protocol, microarrays were run on Clariom D arrays (Affymetrix, Santa Clara, CA) at the Princess Margaret Genomics Centre (Toronto, Canada) to obtain gene expression profiles. Raw microarray data was obtained from previously published work and is available at the Gene Expression Omnibus (GSE127055).[22] A robust multi-array average method in R software (version 2023.09.1) was used to process and normalize the raw microarray data. Brainarray version 25 was used to annotate genes in R software. The ET-1 gene set, which contains 33 related genes was obtained from the molecular signatures database (MSigDB, v3.0). Gene Set Enrichment Analysis (GSEA) (version 4.3.2) was used to determine the enrichment of the ET-1 gene set. Single sample GSEA (ssGSEA) was performed using the GenePattern software (version 2.0) to evaluate the enrichment of the ET-1 gene set within a single sample's expression profile.

Perfusate sample collection & ELISA

Perfusate samples were collected from the LA sampling port on the EVLP system for n = 154 human EVLP cases. Perfusate samples were snap-frozen in liquid nitrogen and transferred to a -80°C freezer for later analysis. Perfusate samples were collected at 60, 90, 110, 120, 130, 150, and 180 minutes of EVLP. ET-1 was measured in all perfusate samples using ELISA (Protein Simple, San Jose, CA, USA) following the Bio-techne User Guide for ELLA (2017). ET-1 levels are reported in pg/mL. The perfusate solution is removed and exchanged during EVLP. A dilution correction calculation was applied to all samples to account for this exchange.

Time series modeling

Seven mathematical models, including linear, quadratic, 4 parameter logistic (PL), 5PL, cubic, quartic, quintic mathematical models, were constructed in R software. Each mathematical model was fit to the ET-1 data for each case. The best fit mathematical model was determined by comparing the median goodness of fit (R) and adjusted R^2 values as well as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) values were used to evaluate model overfitting. Features of the best fit model (coefficients and intercept) were inputted as univariant features into a multiple linear repression (MLR) model to evaluate the predictive value of time series features. Given that ET-1 is a vasoconstrictor and can increase pulmonary vascular resistance (PVR),[24,25] PVR information was added into the MLR model. Similarly, hourly measures of ET-1 and PVR were inputted as univariant features into multiple linear repression (MLR) model to evaluate the predictive value of hourly timepoints.

Integration into a machine learning model – InsighTx

InsighTx, a machine-learning (ML), utilized clinical EVLP data and donor information to predict donor lung suitability of LTx.[23] ET-1 protein levels (at 60, 120, and 180 minutes of EVLP) were integrated into the InsighTx model using Python Programming Language (Python Software (v3.9). The InsighTx model was re-trained in Python using the three new features of ET-1 along with the previously validated features used in the InsighTx model. An XGBoost multiclass classifier was re-trained in order to incorporate the new features (ET-1 hourly data) and account for new model weights. Leave-one-out cross-validation was used and repeated for all cases. To determine feature importance, SHapley Additive exPlanations (SHAP) values were determined.

Statistical analysis

All statistical analyses were performed in GraphPad Prism (version 10.3.0). The selection of a statistical test was determined by the normality of the data. Single genes and ssGSEA scores were analyzed using a paired Wilcoxon test. Pearson's Chi-square test (2-sided) was used to assess the relationship of categorical donor data to the ssGSEA scores and END1 gene expression levels. A Mann-Whitney t-test was used to analyze the relationship of continuous donor data to the ssGSEA scores and EDN1 gene expression levels. DeLong's test for two ROC curves was used to compare the AUROCs.

RESULTS

Significant upregulation of the ET-1 gene set post-EVLP

ssGSEA was used to analyze changes in the ET-1 gene set expression following EVLP using a paired analysis. The ssGSEA results show a significant upregulation of the ET-1 gene set post-EVLP ($p < 0.0001$) (Fig 1 I). GSEA was then used to determine the effect of EVLP on the ET-1 pathway between pre- versus post-EVLP. The enrichment score (ES) for the ET-1 gene set is 0.4242 ($p = 0.17$) (Fig 1 II). This indicated that the ET-1 gene set is upregulated post-EVLP compared to the pre-EVLP expression levels. In addition to the ES, the GSEA software produces a list of the core enriched genes of the ET-1 gene set. The core enriched genes are the genes that contributed most significantly to the enrichment of the gene set. The core enriched genes show a significant upregulation post-EVLP (Figure 1 III). The core enriched genes code for proteins which have various roles in the ET-1 pathway (Fig 1 IV). Interestingly, the EDN1 gene, that codes for the ET-1 protein, was the top core enriched gene.

Clinical relevance of ET-1 gene expression

To assess the clinical significance of the ET-1 pathway, we analyze the relationship of the ssGSEA scores for the ET-1 gene set to donor characteristics. ssGSEA scores were split evenly into a lower (n = 44) and higher (n = 44) expression group. Donor characteristics analyzed were age, BMI, sex, donor types, length of CIT, and smoking status. Pre- and post-EVLP ssGSEA scores were analyzed in relation to various donor characteristics. Additionally, the change in the ssGSEA score from pre- to post-EVLP was analyzed in relation to donor characteristics. Interestingly, donor BMI was significantly lower when the change in the ssGSEA score was higher (26.05±6.42 versus 28.86±7.99, p = 0.039). A significantly higher length of CIT was seen when the change in the ssGSEA score was higher (12.46±2.15 versus 12.04±2.23, p = 0.018).

ET-1 protein (EDN1) was analyzed in relation to various donor characteristics. The same analysis technique was performed, where the EDN1 gene expression level was split evenly into a lower and higher group. There was a significantly higher incidence of donor smoking use in the higher EDN1 expression group pre-EVLP (75.0% versus 47.4%, p = 0.019).

Time-series modeling of ET-1 improves predictive power

Previous unpublished work from our group shows that time series modeling of cytokine biomarkers improves the ability to predict post-transplant outcomes. To assess ET-1's ability to predict donor lung outcomes, we applied time series modeling to ET-1 data. The quartic model was determined to be the best fit for the ET-1 data. Similarly, a linear model provided the best fit for PVR data.

Using mathematical model features in an MLR analysis improved the ability to distinguish declined versus transplanted lungs (AUROC = 79.83 ± 14.46%), compared to hourly measures (AUROC = 65.70 ± 16.95%, p = 0.05691) (Figure 2A). For predicting recipient extubation time (≤72 versus >72 hours), model features significantly improved prediction (AUROC = 94.24 ± 6.32%) over hourly measures (62.42% ± 20.20%) (p = 0.0014) (Figure 2B).

Importance of ET-1 as a biomarker to predict donor lung outcomes

InsighTx is an ML model that utilizes clinical EVLP data, including biochemical, biological (IL-6, IL-8, IL-1β, and IL-10), and physiological data, as well as donor information to predict donor lung suitability for LTx.[23] To assess the importance of ET-1 as a biomarker, we integrated ET-1 hourly timepoints into the InsighTx model and compared the SHAP values between features. A SHAP value over 0 indicates that the model learned new information from that feature, and it was required for the model in making its prediction, whereas a SHAP value equal to zero indicates that the feature was not required. ET-1 shows to be highly ranked when predicting all donor lung outcomes (Figure 3), especially for predicting poor recipient outcomes (Figure 3B).

CONCLUSIONS

This study highlights the significance of ET-1 in EVLP by demonstrating its upregulation post-EVLP and its predictive value in lung transplant outcomes. ssGSEA and GSEA analyses confirmed the upregulation of the ET-1 gene set following EVLP, with core enriched genes, including EDN1, contributing to this effect. Notably, a greater change in ssGSEA scores from pre- to post-EVLP was associated with a lower donor BMI and a longer cold ischemic time. Furthermore, EDN1 expression was significantly associated with donor smoking history, with a higher prevalence of smoking in donors with elevated EDN1 expression pre-EVLP. Time-series modeling further emphasized the value of ET-1, as mathematical model features improved the prediction of lung transplant suitability and post-transplant outcomes compared to hourly measures. Additionally, integrating ET-1 into the InsighTx ML model revealed its importance in predicting donor lung suitability, particularly for poor recipient outcomes. These findings support

ET-1 as a critical biomarker in EVLP and reinforce the potential of advanced modeling techniques in improving lung transplant decision-making.

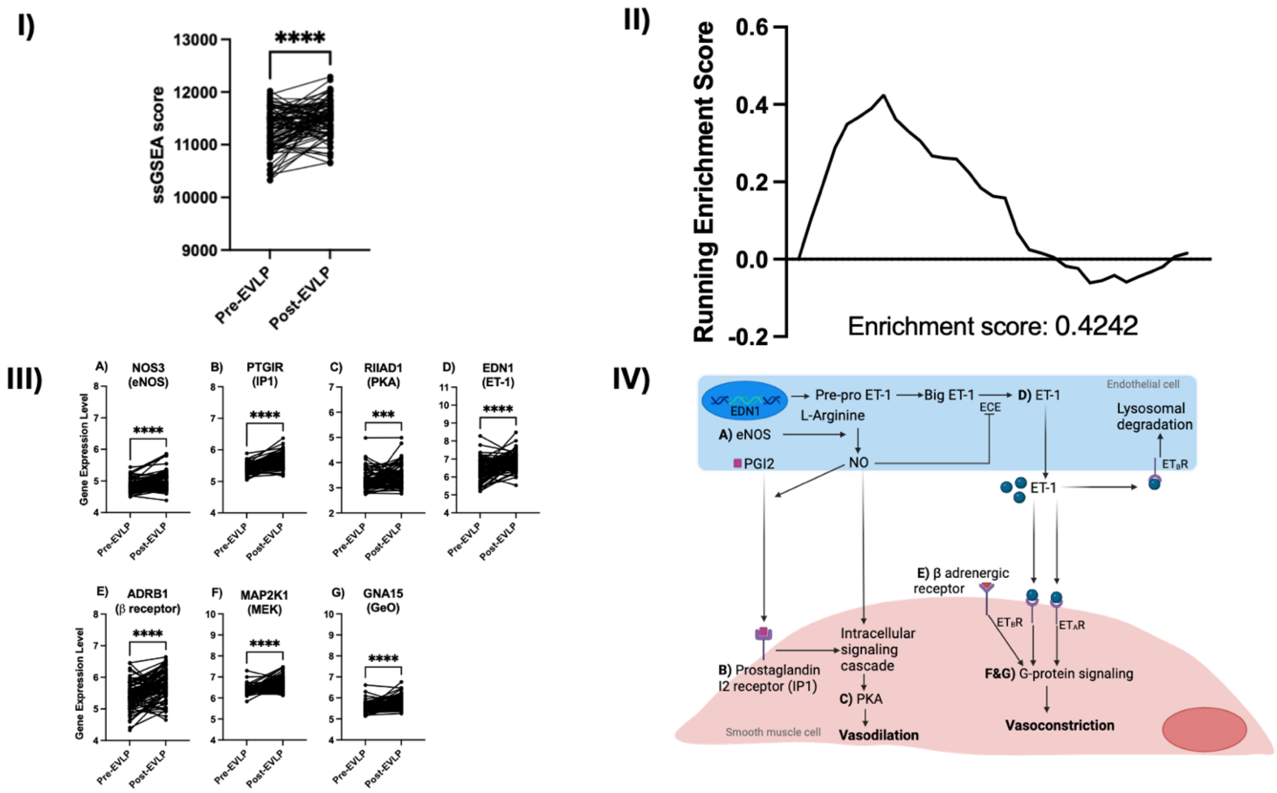


Figure 1. Upregulation of the ET-1 gene set and single genes post-EVLP compared to pre EVLP levels. I) ssGSEA scores of the ET-1 gene set pre-EVLP compared to post EVLP. II) Gene set enrichment analysis of 33 genes in the ET-1 gene set comparing the expression from pre-EVLP to post-EVLP in human lung tissue. III) Gene expression levels at pre- and post-EVLP of A) NOS3, B) PTGIR, C) RIIAD1, D) EDN1, E) ADRB1, F) MAP2K1, and G) GNA15 and IV) their relation to the ET-1 pathway. Statistical analysis was done using a paired nonparametric t-test (Wilcoxon test). Statistical significance was defined as $p < 0.05$, n= 88. *** p <0.001, **** p <0.0001.

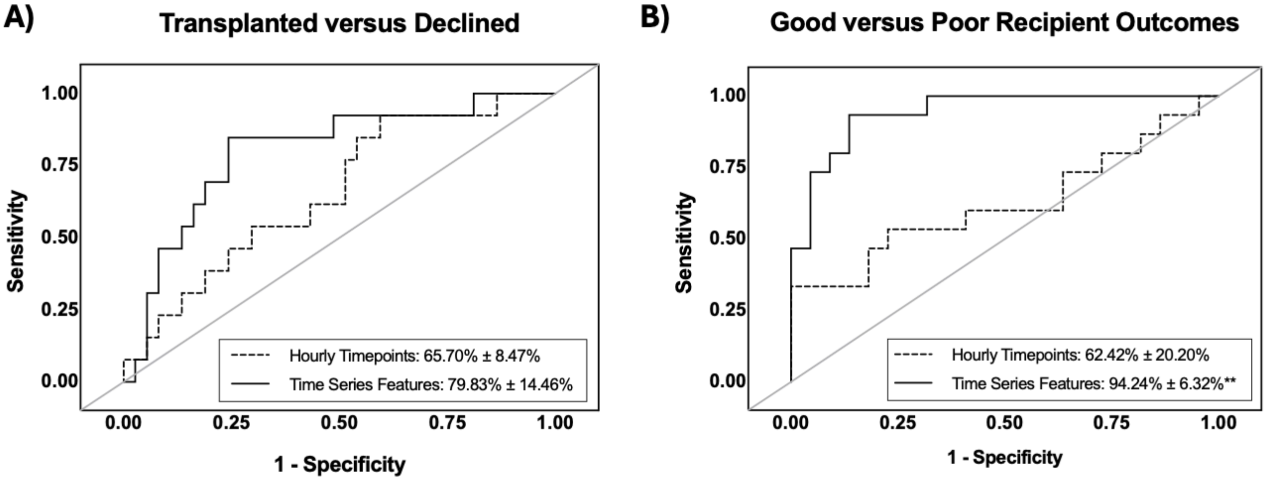


Figure 2. AUROC curves for predicting A) transplant versus declined lungs and B) good versus poor recipient outcomes. AUROC curves using hourly timepoints of ET-1 and PVR and using time series features of ET-1 and PVR. Statistical analysis was done using the DeLong's test for two ROC curves. Statistical significance was defined as $p < 0.05$. ** p-value < 0.01.

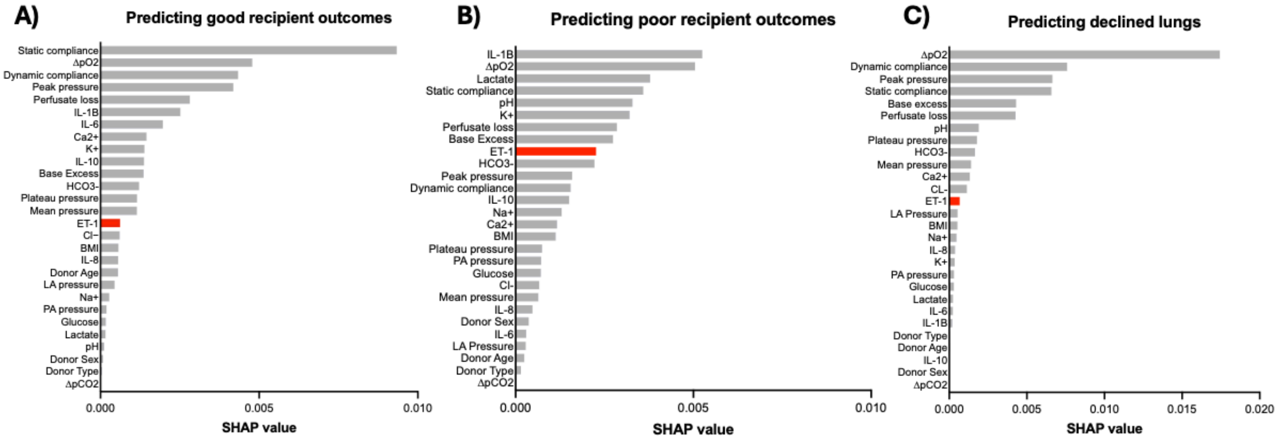


Figure 3. SHAP value plots for predicting donor lung outcomes. Feature importance for predicting A) good recipient outcomes, B) poor recipient outcomes, and C) declined lungs.

References

1. WHO, Transplantation Society (TTS) & Organizatión Nacional de Transplantes (ONT). Third WHO Global Consultation on Organ Donation and Transplantation: striving to achieve self-sufficiency, March 23–25, 2010, Madrid, Spain. *Transplantation* **91 Suppl 11**, S27-8 (2011).

2. Israni, A. K., Zaun, D. A., Rosendale, J. D., Snyder, J. J. & Kasiske, B. L. OPTN/SRTR 2013 Annual Data Report: Deceased Organ Donation. *American Journal of Transplantation* **15**, 1–13 (2015).

3. Snell, G. I. *et al.* Report of the ISHLT Working Group on Primary Lung Graft Dysfunction, part I: Definition and grading—A 2016 Consensus Group statement of the International Society for Heart and Lung Transplantation. *Journal of Heart and Lung Transplantation* vol. 36 1097–1103 Preprint at https://doi.org/10.1016/j.healun.2017.07.021 (2017).

4. Steen, S. *et al.* Transplantation of lungs from a non-heart-beating donor. *The Lancet* **357**, 825–829 (2001).

5. M, M., Attawar, S., BN, M., Tisekar, O. & Mohandas, A. Ex vivo lung perfusion and the Organ Care System: a review. *Clinical Transplantation and Research* **38**, 23–36 (2024).

6. Chao, B. T. *et al.* A radiographic score for human donor lungs on ex vivo lung perfusion predicts transplant outcomes. *Journal of Heart and Lung Transplantation* **43**, 797–805 (2024).

7. Chao, B. T. *et al.* Improving prognostic accuracy in lung transplantation using unique features of isolated human lung radiographs. *NPJ Digit Med* **7**, (2024).

8. Koike, T. *et al.* Kinetics of lactate metabolism during acellular normothermic ex vivo lung perfusion. *Journal of Heart and Lung Transplantation* **30**, 1312–1319 (2011).

9. Di Nardo, M. *et al.* Predicting donor lung acceptance for transplant during ex vivo lung perfusion: The EX vivo lung PerfusIon pREdiction (EXPIRE). *American Journal of Transplantation* **21**, 3704–3713 (2021).

10. Watanabe, T., Cypel, M. & Keshavjee, S. Ex vivo lung perfusion. *Journal of Thoracic Disease* vol. 13 6602–6617 Preprint at https://doi.org/10.21037/jtd-2021-23 (2021).

11. Sage, A. T. *et al.* Prediction of donor related lung injury in clinical lung transplantation using a validated ex vivo lung perfusion inflammation score. *Journal of Heart and Lung Transplantation* **40**, 687–695 (2021).

12. Andreasson, A. S. I. *et al.* The role of interleukin-1β as a predictive biomarker and potential therapeutic target during clinical ex vivo lung perfusion. *The Journal of Heart and Lung Transplantation* **36**, 985–995 (2017).

13. Major, T. *et al.* Pro-IL-1β Is an Early Prognostic Indicator of Severe Donor Lung Injury During Ex Vivo Lung Perfusion. *Transplantation* **105**, 768–774 (2021).

14. Machuca, T. N. *et al.* Protein Expression Profiling Predicts Graft Performance in Clinical Ex Vivo Lung Perfusion. *Ann Surg* **261**, 591–597 (2015).

15. Hashimoto, K. *et al.* Higher M30 and high mobility group box 1 protein levels in ex vivo lung perfusate are associated with primary graft dysfunction after human lung transplantation. *J Heart Lung Transplant* (2017) doi:10.1016/j.healun.2017.06.005.

16. Caldarone, L. *et al.* Neutrophil extracellular traps in *ex vivo* lung perfusion perfusate predict the clinical outcome of lung transplant recipients. *European Respiratory Journal* **53**, 1801736 (2019).

17. Kanou, T. *et al.* Cell-free DNA in human ex vivo lung perfusate as a potential biomarker to predict the risk of primary graft dysfunction in lung transplantation. *J Thorac Cardiovasc Surg* **162**, 490-499.e2 (2021).

18. Machuca, T. N. *et al.* The role of the endothelin-1 pathway as a biomarker for donor lung assessment in clinical ex vivo lung perfusion. *Journal of Heart and Lung Transplantation* **34**, 849–857 (2015).

19. Hashimoto, K. *et al.* Soluble Adhesion Molecules During Ex Vivo Lung Perfusion Are Associated With Posttransplant Primary Graft Dysfunction. *American Journal of Transplantation* **17**, 1396–1404 (2017).

20. Sladden, T. M. *et al.* Endothelial Glycocalyx Shedding Occurs during Ex Vivo Lung Perfusion: A Pilot Study. *J Transplant* **2019**, 1–12 (2019).

21. Salama, M. *et al.* Concomitant endothelin-1 overexpression in lung transplant donors and recipients predicts primary graft dysfunction. *American Journal of Transplantation* **10**, 628–636 (2010).

22. Wong, A. *et al.* Potential therapeutic targets for lung repair during human ex vivo lung perfusion. (2019) doi:10.1183/13993003.02222.

23. Sage, A. T. *et al.* A machine-learning approach to human ex vivo lung perfusion predicts transplantation outcomes and promotes organ utilization. *Nat Commun* **14**, (2023).

24. Giaid, A. *et al.* Expression of endothelin-1 in the lungs of patients with pulmonary hypertension. *N Engl J Med* **328**, 1732–9 (1993).

25. Shao, D., Park, J. E. S. & Wort, S. J. The role of endothelin-1 in the pathogenesis of pulmonary arterial hypertension. *Pharmacological Research* vol. 63 504–511 Preprint at https://doi.org/10.1016/j.phrs.2011.03.003 (2011).

USING MACHINE LEARNING TO PREDICT OUTCOMES FOLLOWING OPEN ABDOMINAL AORTIC ANEURYSM REPAIR

Ben Li MD (SSTP)[1,2,3,4], Badr Aljabri[5], Raj Verma[6], Derek Beaton[7], Naomi Eisenberg[8], Douglas S Lee[9,10,11], Duminda N Wijeysundera[10,11,12,13], Thomas L Forbes[1,3,8], Ori D Rotstein [1,3,13,14], Charles de Mestral[1,2,10,11,13], Muhammad Mamdani[3,4,7,10,11,13,15], Graham Roche-Nagle[1,8], Mohammed Al-Omran[1,2,3,4,13,16]

[1]Department of Surgery, University of Toronto, Canada
[2]Division of Vascular Surgery, St. Michael's Hospital, Unity Health Toronto, Canada
[3]Institute of Medical Science, University of Toronto, Canada
[4]Temerty Centre for Artificial Intelligence Research and Education in Medicine (T-CAIREM), University of Toronto, Canada
[5]Department of Surgery, King Saud University, Saudi Arabia
[6]School of Medicine, Royal College of Surgeons in Ireland, University of Medicine and Health Sciences, Ireland
[7]Data Science & Advanced Analytics, Unity Health Toronto, University of Toronto, Canada
[8]Division of Vascular Surgery, Peter Munk Cardiac Centre, University Health Network, Canada
[9]Division of Cardiology, Peter Munk Cardiac Centre, University Health Network, Canada
[10]Institute of Health Policy, Management and Evaluation, University of Toronto, Canada
[11]ICES, University of Toronto, Canada
[12]Department of Anesthesia, St. Michael's Hospital, Unity Health Toronto, Canada
[13]Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Canada
[14]Division of General Surgery, St. Michael's Hospital, Unity Health Toronto, Canada
[15]Leslie Dan Faculty of Pharmacy, University of Toronto, Canada
[16]Department of Surgery, King Faisal Specialist Hospital and Research Center, Saudi Arabia

INTRODUCTION

Elective open surgical repair is indicated for abdominal aortic aneurysms (AAA) with diameters above 5.0 cm in women and 5.5 cm in men[1]; however, the procedure carries a high rate of complications[2]. Deery and colleagues (2016) demonstrated that major adverse events occur in up to 20% of patients undergoing open AAA repair and the risk is heightened for complex aneurysms[3]. As a result, the Society for Vascular Surgery (SVS) and European Society for Vascular Surgery (ESVS) AAA guidelines recommend careful surgical risk assessment when considering patients for intervention[4,5].

There are currently no standardized tools to predict complications following open AAA repair. A systematic review of 13 risk prediction models demonstrated significant methodological limitations and variable performance across different populations[6]. Furthermore, tools such as the SVS Vascular Quality Initiative (VQI) Cardiac Risk Index (CRI)[7] and American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP)[8] online surgical risk calculators use modelling techniques that require manual input of clinical variables, which deters routine use in busy medical settings[9]. Therefore, there is an important need to develop better and more practical surgical risk prediction tools for patients undergoing open AAA repair.

Machine learning (ML) is a rapidly advancing technology that allows computers to learn from data and make predictions[10]. This field has been driven by the explosion of electronic medical record data combined with increasing computational power[11]. The advantage of newer ML techniques over traditional statistical methods is that they can better model complex, multicollinear relationships between covariates and outcomes[12], which is common in health care data[13]. The VQI database is a large, multicentre vascular registry that contains highly granular and procedure-specific variables, which is ideal for building robust ML models[14]. In this study, we used VQI data to develop ML models that can accurately predict outcomes following elective open AAA repair using pre-operative data. We hypothesized that ML algorithms could achieve better predictive performance compared to traditional statistical models such as logistic regression.

METHODS

Study approval

The Research Advisory Council of the SVS Patient Safety Organization (PSO) approved this project and provided the blinded data.

Design

We conducted a ML-based prognostic study and reported our findings based on the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence (TRIPOD + AI) statement[15].

Dataset

The Vascular Quality Initiative (VQI) database is a clinical registry maintained by the SVS PSO with the goal of improving the delivery of vascular care[14]. Over 1,000 academic and community hospitals worldwide prospectively submit demographic, clinical, and outcomes data on consecutive eligible vascular surgery patients, including information from their initial hospitalization up to 9-21 months of follow-up[16]. Annual audits with comparison to hospital claims are performed to ensure accuracy of the submitted information[17].

Patient Cohort

All patients who underwent elective open AAA repair between January 1, 2003 and January 4, 2023 in the VQI database were included. Patients with ruptured or symptomatic AAA were excluded.

Features

Given the unique advantage of ML techniques in handling large numbers of input features, all available pre-operative variables in the VQI database (n = 52) were used to maximize model performance. Demographic variables included age, sex, body mass index, race, ethnicity, insurance status, rural residence, median area deprivation index (ADI), and transfer status. Comorbidities included smoking status, hypertension, diabetes, family history of AAA, coronary artery disease (CAD), congestive heart failure (CHF), chronic obstructive pulmonary disease (COPD), end stage renal disease (ESRD) requiring dialysis, and American Association of Anesthesiologists (ASA) classification. Other variables included previous vascular procedures, functional status, investigations (hemoglobin, creatinine, cardiac stress test results, and ejection fraction), medications, anatomic characteristics, and concurrent procedures.

Outcomes

The primary outcome was in-hospital major adverse cardiovascular event (MACE), defined as a composite of myocardial infarction (MI), stroke, or death. MI was defined as a combination of clinical, electrocardiogram, and biomarker evidence of acute myocardial ischemia. Stroke was defined as acute, focal neurological deficits persisting for > 24 hours with clinical and/or imaging evidence of vascular injury to the central nervous system. Death was defined as all-cause mortality. In-hospital MACE was chosen as the primary outcome because these complications are generally directly related to open AAA repair and have an important impact on morbidity and mortality[18,19]. Secondary outcomes were 1-year mortality and 1-year reintervention related to the incision, graft, bowel, or lower extremity ischemia.

Model development

Six ML models were trained to predict primary and secondary outcomes: Extreme Gradient Boosting (XGBoost), random forest, Naïve Bayes classifier, radial basis function (RBF) support vector machine (SVM), multilayer perceptron (MLP) artificial neural network (ANN), and logistic regression. These ML algorithms are widely used in the literature and demonstrate the best performance for predicting categorical surgical outcomes[20–22].

Our data were randomly split into training (70%) and test (30%) sets[23]. Ten-fold cross-validation and grid search were performed on the training set to find optimal model hyperparameters[24,25]. To improve class balance, Random Over-Sample Examples (ROSE) was applied to the training set[26]. The models were then evaluated on unseen data in the test set and ranked based on discriminatory metrics, primarily area under the receiver operating characteristic curve (AUROC). The best performing model was XGBoost, which had the following optimized hyperparameters on our dataset: number of rounds = 150, maximum tree depth = 3, learning rate = 0.3, gamma = 0, column sample by tree = 0.6, minimum child weight = 1, subsample = 0.9.

Statistical analysis

Baseline characteristics were summarized as means (standard deviation), medians (interquartile range), or numbers (proportion). Differences between groups were assessed using independent t-tests for continuous variables or chi-square tests for categorical variables. Bonferroni correction was used to set statistical significance to account for multiple comparisons.

The primary metric for assessing model performance was AUROC (95% CI). Secondary performance metrics were accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). To further assess predictive performance, we plotted calibration curves and calculated Brier scores[27]. In the final model, feature importance was determined by ranking the top 10 predictors based on variable importance scores (gain)[28]. To assess model robustness on various populations, we performed subgroup analysis of predictive performance based on age, sex, race, ethnicity, rurality, socioeconomic status, proximal clamp site, prior aortic surgery, and concomitant procedures.

Based on a validated sample size calculator for clinical prediction models, to achieve a minimum AUROC of 0.8 with an outcome rate of ~5% and 52 pre-operative features, the minimum sample size required is 7,274 patients with 364 events[29]. Our cohort of 12,027 patients with 630 primary events met this sample size requirement. There were less than 5% missing data for variables of interest; therefore, complete-case analysis was applied[30]. Patients lost to follow-up for 1-year outcomes were censored. All analyses were performed in R version 4.2.1[31].

RESULTS

Patients, events, and follow-up

From an initial cohort of 16,918 patients who underwent open AAA repair in the VQI database from 2003-2023, we excluded 2,910 patients for ruptured AAA and 1,981 patients for symptomatic AAA. Overall, we included 12,027 patients, and 630 (5.2%) had the primary outcome of in-hospital MACE. The individual components of MACE occurred in the following distribution: MI (n=227 [1.9%]), stroke (n=87 [0.7%]), and death (n=411 [3.4%]). For secondary outcomes, 1-year mortality occurred in 828 (6.9%) patients and 1-year re-intervention occurred in 600 (5.0%) patients. Mean follow-up was 15.3 (SD 7.7) months.

Pre-operative demographic and clinical characteristics

Compared to patients without a primary outcome, those who developed in-hospital MACE were older (mean age 73.1 [SD 7.0] vs. 69.0 [SD 8.7], $p < 0.001$) and more likely to be female (32.4% vs. 25.5%, $p < 0.001$). They were also more likely to have hypertension, diabetes, CAD, CHF, COPD, ESRD requiring dialysis, and an ASA class ≥4. Functionally, patients with an event were more likely to reside in nursing homes, require assistance for ambulation, or were wheelchair-dependent/bedridden. For investigations, patients with in-hospital MACE had a higher mean creatinine level and were more likely to have a positive cardiac stress test and ejection fraction below 50%. Anatomically, patients with a primary outcome had a larger mean AAA diameter with a greater proportion requiring a suprarenal or supraceliac clamp, distal anastomosis to the common femoral artery, and concurrent renal/infrainguinal bypass or other abdominal surgical procedure (Table 1).

Model performance

Of the 6 ML models evaluated on test set data for predicting in-hospital MACE following open AAA repair, XGBoost had the best performance with an AUROC (95% CI) of 0.93 (0.92-0.94) compared to random forest [0.89 (0.87-0.90)], RBF SVM [0.85 (0.83-0.86)], Naïve Bayes [0.82 (0.80-0.83)], MLP ANN [0.80 (0.78-0.82)], and logistic regression [0.71 (0.70-0.73)]. The other performance metrics of XGBoost were the following: accuracy 0.86 (95% CI 0.84-0.87), sensitivity 0.84, specificity 0.87, PPV 0.88, and NPV 0.83 (Table 2). For predicting 1-year mortality and re-intervention, XGBoost achieved AUROC's (95% CI) of 0.93 (0.92-0.94) and 0.84 (0.82-0.85), respectively.

The ROC curve for predicting in-hospital MACE using XGBoost is demonstrated in Figure 1. Our model achieved good calibration with a Brier score of 0.05, indicating excellent agreement between predicted and observed evented probabilities. The top 10 predictors of in-hospital MACE in our XGBoost model were the following: 1) CAD, 2) ASA class, 3) proximal clamp site, 4) CHF, 5) prior carotid revascularization, 6) pre-operative ambulation status, 7) COPD, 8) older age, 9) concurrent renal bypass, and 10) pre-operative creatinine. Model performance remained robust on all subgroup analyses of specific demographic/clinical populations, with AUROC's ranging from 0.92-0.94 and no significant differences between majority and minority groups.

CONCLUSIONS

In this study, we used a large clinical registry to develop automated, explainable, and robust ML models that predict in-hospital and 1-year outcomes following open AAA repair with excellent performance using pre-operative data (AUROC's ≥ 0.90). Given that our ML algorithms perform better than existing tools[6] and logistic regression, they have potential for important utility in the peri-operative management of patients being considered for open AAA repair to mitigate adverse outcomes. Prospective validation of our prediction models is warranted.

ACKNOWLEDGEMENTS

None.

SOURCES OF FUNDING

This research was funded by the University of Toronto Surgeon Scientist Training Program, Canadian Institutes of Health Research, Ontario Ministry of Health, and PSI Foundation (BL). The funding sources did not play a role in the design or conduct of the research.

DISCLOSURES

The authors have no conflicts of interest.

CODE AVAILABILITY STATEMENT

The complete code used for model development and evaluation in this project is publicly available on GitHub: https://github.com/benli12345/OAAA-ML-VQI.

DATA AVAILABILITY STATEMENT

The data used for this study comes from the Vascular Quality Initiative Database, which is maintained by the Society for Vascular Surgery Patient Safety Organization. Access to and use of the data requires approval through an application process available at https://www.vqi.org/data-analysis/.

FULL-TEXT PUBLICATION

Li B, Aljabri B, Verma R, Beaton D, Eisenberg N, Lee DS, Wijeysundera DN, Forbes TL, Rotstein OD, de Mestral C, Mamdani M, Roche-Nagle G, Al-Omran M. Using machine learning to predict outcomes following open abdominal aortic aneurysm repair. Journal of Vascular Surgery. 2023 Dec 1;78(6):1426-1438.e6. https://www.jvascsurg.org/article/S0741-5214(23)01935-3/fulltext

REFERENCES

1. Shaw PM, Loree J, Gibbons RC. Abdominal Aortic Aneurysm. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cited 2022 Dec 26]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK470237/

2. Menard MT, Chew DKW, Chan RK, Conte MS, Donaldson MC, Mannick JA, et al. Outcome in patients at high risk after open surgical repair of abdominal aortic aneurysm. J Vasc Surg. 2003 Feb;37(2):285–92.

3. Deery SE, Lancaster RT, Baril DT, Indes JE, Bertges DJ, Conrad MF, et al. Contemporary outcomes of open complex abdominal aortic aneurysm repair. J Vasc Surg. 2016 May;63(5):1195–200.

4. Wanhainen A, Verzini F, Van Herzeele I, Allaire E, Bown M, Cohnert T, et al. Editor's Choice - European Society for Vascular Surgery (ESVS) 2019 Clinical Practice Guidelines on the Management of Abdominal Aorto-iliac Artery Aneurysms. Eur J Vasc Endovasc Surg Off J Eur Soc Vasc Surg. 2019 Jan;57(1):8–93.

5. Chaikof EL, Dalman RL, Eskandari MK, Jackson BM, Lee WA, Mansour MA, et al. The Society for Vascular Surgery practice guidelines on the care of patients with an abdominal aortic aneurysm. J Vasc Surg. 2018 Jan 1;67(1):2-77.e2.

6. Lijftogt N, Luijnenburg TWF, Vahl AC, Wilschut ED, Leijdekkers VJ, Fiocco MF, et al. Systematic review of mortality risk prediction models in the era of endovascular abdominal aortic aneurysm surgery. Br J Surg. 2017 Jul;104(8):964–76.

7. Bertges DJ, Neal D, Schanzer A, Scali ST, Goodney PP, Eldrup-Jorgensen J, et al. The Vascular Quality Initiative Cardiac Risk Index for prediction of myocardial infarction after vascular surgery. J Vasc Surg. 2016 Nov;64(5):1411-1421.e4.

8. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, et al. Development and Evaluation of the Universal ACS NSQIP Surgical Risk Calculator: A Decision Aid and Informed Consent Tool for Patients and Surgeons. J Am Coll Surg. 2013 Nov;217(5):833–42.

9. Sharma V, Ali I, van der Veer S, Martin G, Ainsworth J, Augustine T. Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. BMJ Health Care Inform. 2021 Feb 19;28(1):e100253.

10. Baştanlar Y, Özuysal M. Introduction to machine learning. Methods Mol Biol. 2014;1107:105–28.

11. Shah P, Kendall F, Khozin S, Goosen R, Hu J, Laramie J, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit Med. 2019;2:69.

12. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019 May;20(5):e262–73.

13. Liew BXW, Kovacs FM, Rügamer D, Royuela A. Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain. Eur Spine J

Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc. 2022 Aug;31(8):2082–91.

14. Society for Vascular Surgery Vascular Quality Initiative (VQI) [Internet]. [cited 2022 Jul 11]. Available from: https://www.vqi.org/

15. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378.

16. Vascular Quality Initiative [Internet]. [cited 2023 Jun 16]. Available from: https://www.vqi.org/

17. Cronenwett JL, Kraiss LW, Cambria RP. The Society for Vascular Surgery Vascular Quality Initiative. J Vasc Surg. 2012 May;55(5):1529–37.

18. Brown LC, Thompson SG, Greenhalgh RM, Powell JT, Endovascular Aneurysm Repair trial participants. Incidence of cardiovascular events and death after open or endovascular repair of abdominal aortic aneurysm in the randomized EVAR trial 1. Br J Surg. 2011 Jul;98(7):935–42.

19. Columbo JA, Demsas F, Wanken ZJ, Suckow BD, Beach JM, Henkin S, et al. Stress testing before abdominal aortic aneurysm repair does not lead to a reduction in perioperative cardiac events. J Vasc Surg. 2021 Sep;74(3):694–700.

20. Elfanagely O, Toyoda Y, Othman S, Mellia JA, Basta M, Liu T, et al. Machine Learning and Surgical Outcomes Prediction: A Systematic Review. J Surg Res. 2021 Apr 10;264:346–61.

21. Bektaş M, Tuynman JB, Costa Pereira J, Burchell GL, van der Peet DL. Machine Learning Algorithms for Predicting Surgical Outcomes after Colorectal Surgery: A Systematic Review. World J Surg. 2022 Sep 15;

22. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. World Neurosurg. 2018 Jan;109:476-486.e1.

23. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med Genomics. 2011 Apr 8;4:31.

24. Jung Y, Hu J. A K-fold Averaging Cross-validation Procedure. J Nonparametric Stat. 2015;27(2):167–79.

25. Adnan M, Alarood AAS, Uddin MI, Ur Rehman I. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. PeerJ Comput Sci. 2022;8:e803.

26. Wibowo P, Fatichah C. Pruning-based oversampling technique with smoothed bootstrap resampling for imbalanced clinical dataset of Covid-19. J King Saud Univ - Comput Inf Sci. 2022 Oct;34(9):7830–9.

27. Redelmeier DA, Bloch DA, Hickam DH. Assessing predictive accuracy: how to compare Brier scores. J Clin Epidemiol. 1991;44(11):1141–6.

28. Loh WY, Zhou P. Variable Importance Scores. J Data Sci. 2021 Sep 16;19(4):569–92.

29. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020 Mar 18;m441.

30. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol. 2019 Aug;48(4):1294–304.

31. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Table 1. Pre-operative demographic and clinical characteristics of patients undergoing open abdominal aortic aneurysm repair with and without in-hospital major adverse cardiovascular events

	Absence of in-hospital MACE (n = 11,397)	**Presence of in-hospital MACE (n = 630)**	**P**
Demographics			
Age, years, mean (SD)	69.0 (8.7)	73.1 (7.0)	< 0.001
Female	2,901 (25.5)	204 (32.4)	< 0.001
BMI, kg/m^2, mean (SD)	27.5 (5.5)	26.9 (6.3)	0.009
Race			
American Indian or Alaskan Native	33 (0.3)	2 (0.3)	0.008
Asian	156 (1.4)	14 (2.2)	
Black	484 (4.3)	26 (4.1)	
Native Hawaiian or other Pacific Islander	11 (0.1)	4 (0.6)	
White	10,025 (88.0)	550 (87.3)	
More than 1 race	15 (0.1)	1 (0.2)	
Unknown/other	673 (5.9)	33 (5.2)	
Hispanic ethnicity	238 (2.1)	12 (1.9)	0.86
Insurance status			
Medicare	4,386 (38.5)	277 (44.0)	0.008
Medicaid	358 (3.1)	12 (1.9)	
Commercial	3,291 (28.9)	139 (22.1)	
Medicare Advantage	485 (4.3)	29 (4.6)	
Military/Veterans Affairs	180 (1.6)	10 (1.6)	
Non-US Insurance	672 (5.9)	38 (6.0)	
Self-pay (uninsured)	92 (0.8)	6 (1.0)	
Unknown/other	1,933 (17.0)	119 (18.9)	
Rural residence	709 (6.2)	49 (7.8)	0.12
Area Deprivation Index percentile, median (IQR)	54 (39 – 72)	54 (40 – 72)	0.74
Transfer status			
From another hospital	391 (3.4)	16 (2.5)	0.24
From rehabilitation unit	15 (0.1)	2 (0.3)	
Comorbidities			
Smoking status			
Never	1,068 (9.4)	48 (7.6)	0.31
Prior	5,627 (49.4)	322 (51.1)	
Current	4,702 (41.3)	260 (41.3)	
Hypertension	9,583 (84.1)	574 (91.1)	< 0.001
Diabetes	1,886 (16.5)	142 (22.5)	0.002
Family history of AAA	1,291 (11.3)	59 (9.4)	0.15

	Absence of in-hospital MACE (n = 11,397)	**Presence of in-hospital MACE (n = 630)**	**P**
Coronary artery disease	4,149 (36.4)	317 (50.3)	< 0.001
Prior open or endovascular coronary revascularization	3,388 (29.8)	246 (39.0)	< 0.001
Congestive heart failure	853 (7.5)	94 (14.9)	< 0.001
Chronic obstructive pulmonary disease			
Not treated	1,399 (12.3)	103 (16.3)	< 0.001
On medications	2,049 (18.0)	169 (26.8)	
On home oxygen	236 (2.1)	24 (3.8)	
Dialysis	59 (0.5)	9 (1.4)	0.01
ASA Class			
1	43 (0.4)	2 (0.3)	< 0.001
2	483 (4.2)	13 (2.1)	
3	7,406 (65.0)	344 (54.6)	
4	3,448 (30.3)	270 (42.9)	
5	17 (0.1)	1 (0.2)	
Previous procedures			
Prior aortic surgery			
Infrarenal open AAA repair	250 (2.2)	23 (3.7)	0.07
Suprarenal open AAA repair	105 (0.9)	5 (0.8)	
Aortic bypass	48 (0.4)	4 (0.6)	
Endovascular AAA repair	220 (1.9)	12 (1.9)	
Aortic endarterectomy or other	702 (6.2)	50 (7.9)	
Prior extracranial aneurysm repair	1,215 (10.7)	84 (13.3)	0.04
Prior carotid endarterectomy or stent	583 (5.1)	86 (13.7)	< 0.001
Prior bypass for peripheral artery disease	463 (4.1)	48 (7.6)	< 0.001
Prior endovascular intervention for peripheral artery disease	707 (6.2)	84 (13.3)	< 0.001
Prior major amputation	53 (0.5)	2 (0.3)	0.87
Functional status			
Living status			
Home	11,329 (99.4)	623 (98.9)	0.06
Nursing home	54 (0.5)	7 (1.1)	
Homeless	14 (0.1)	0	
Pre-operative ambulatory status			
Independent	10,797 (94.7)	572 (90.8)	< 0.001
With assistance	40 (4.7)	49 (7.8)	
Wheelchair-dependent	55 (0.5)	8 (1.3)	
Bedridden	5 (0.04)	1 (0.2)	

	Absence of in-hospital MACE (n = 11,397)	Presence of in-hospital MACE (n = 630)	P
Investigations			
Hemoglobin, g/L, mean (SD)	137.0 (17.9)	132.0 (19.0)	< 0.001
Creatinine, umol/L, mean (SD)	95.8 (37.9)	108.0 (53.7)	< 0.001
Cardiac stress test			
Not done	4,912 (43.1)	270 (42.9)	0.02
Normal	5,211 (45.7)	265 (42.1)	
Positive for ischemia	716 (6.3)	49 (7.8)	
Positive for infarction	423 (3.7)	33 (5.2)	
Positive for ischemia and infarction	135 (1.2)	13 (2.1)	
Ejection fraction			
< 30%	108 (0.9)	9 (1.4)	< 0.001
30-50%	1,083 (9.5)	87 (13.8)	
> 50%	6,987 (61.3)	397 (63.0)	
Not done	2,430 (21.3)	98 (15.6)	
Unknown	789 (6.9)	39 (6.2)	
Medications			
Acetylsalicylic acid	7,395 (64.9)	438 (69.5)	0.02
P2Y12 antagonist	936 (8.2)	97 (15.4)	< 0.001
Statin	8,320 (73.0)	474 (75.2)	0.24
Beta blocker	6,915 (60.7)	432 (68.6)	< 0.001
ACE-I/ARB	4,453 (39.1)	282 (44.8)	0.005
Anticoagulant	905 (7.9)	63 (10.0)	0.08
Anatomic characteristics and procedural planning variables			
Maximum AAA diameter, cm, mean (SD)	6.0 (1.6)	6.2 (1.5)	0.001
Concomitant iliac artery aneurysm			
Unilateral	1,429 (12.5)	63 (10.0)	0.15
Bilateral	1,842 (16.2)	100 (15.9)	
Surgical exposure			
Transperitoneal	8,185 (71.8)	446 (70.8)	0.86
Retroperitoneal	3,124 (27.4)	179 (28.4)	
Not reported	88 (0.8)	5 (0.8)	
Proximal graft diameter, mm, median (IQR)	18 (16 – 20)	18 (16 – 20)	0.91
Proximal clamp site			
Infrarenal	6,287 (55.2)	287 (45.6)	< 0.001
Above 1 renal artery	1,610 (14.1)	90 (14.3)	
Above both renal arteries	2,595 (22.8)	181 (28.7)	
Supraceliac	711 (6.2)	63 (10.0)	

	Absence of in-hospital MACE (n = 11,397)	**Presence of in-hospital MACE (n = 630)**	**P**
Not reported	194 (1.7)	9 (1.4)	
Distal anastomosis			
Aorta	4,566 (40.1)	219 (34.8)	< 0.001
Common iliac artery	4,421 (38.8)	230 (36.5)	
External iliac artery	998 (8.8)	47 (7.5)	
Common femoral artery	1,238 (10.9)	115 (18.3)	
Not reported	174 (1.5)	19 (3.0)	
Concurrent procedures			
Renal bypass	542 (4.8)	51 (8.1)	< 0.001
Infrainguinal bypass	230 (2.0)	29 (4.6)	< 0.001
Other abdominal procedure	1,036 (9.1)	86 (13.7)	< 0.001

Values are reported as No. (%) unless otherwise indicated. Abbreviations: MACE (major adverse cardiovascular event), AAA (abdominal aortic aneurysm), BMI (body mass index), ACE-I (angiotensin converting enzyme inhibitor), ARB (angiotensin II receptor blocker), ASA (American Association of Anesthesiologists), SD (standard deviation), IQR (interquartile range).

Table 2. Model performance on test set data for predicting in-hospital major adverse cardiovascular events following open abdominal aortic aneurysm repair using pre-operative features

	AUROC (95% CI)	**Accuracy (95% CI)**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
XGBoost	0.93 (0.92 – 0.94)	0.86 (0.84 – 0.87)	0.84	0.87	0.88	0.83
Random forest	0.89 (0.87 – 0.90)	0.80 (0.79 – 0.81)	0.82	0.78	0.77	0.83
RBF SVM	0.85 (0.83 – 0.86)	0.76 (0.75 – 0.78)	0.75	0.78	0.80	0.73
Naïve Bayes	0.82 (0.80 – 0.83)	0.82 (0.80 – 0.83)	0.80	0.84	0.85	0.78
MLP ANN	0.80 (0.78 – 0.82)	0.79 (0.77 – 0.81)	0.77	0.83	0.86	0.72
Logistic regression	0.71 (0.70 – 0.73)	0.62 (0.60 – 0.64)	0.59	0.73	0.66	0.58

Abbreviations: XGBoost (Extreme Gradient Boosting), AUROC (area under the receiver operating characteristic curve), CI (confidence interval), PPV (positive predictive value), NPV (negative predictive value), RBF SVM (radial basis function support vector machine), MLP ANN (multilayer perceptron artificial neural network).

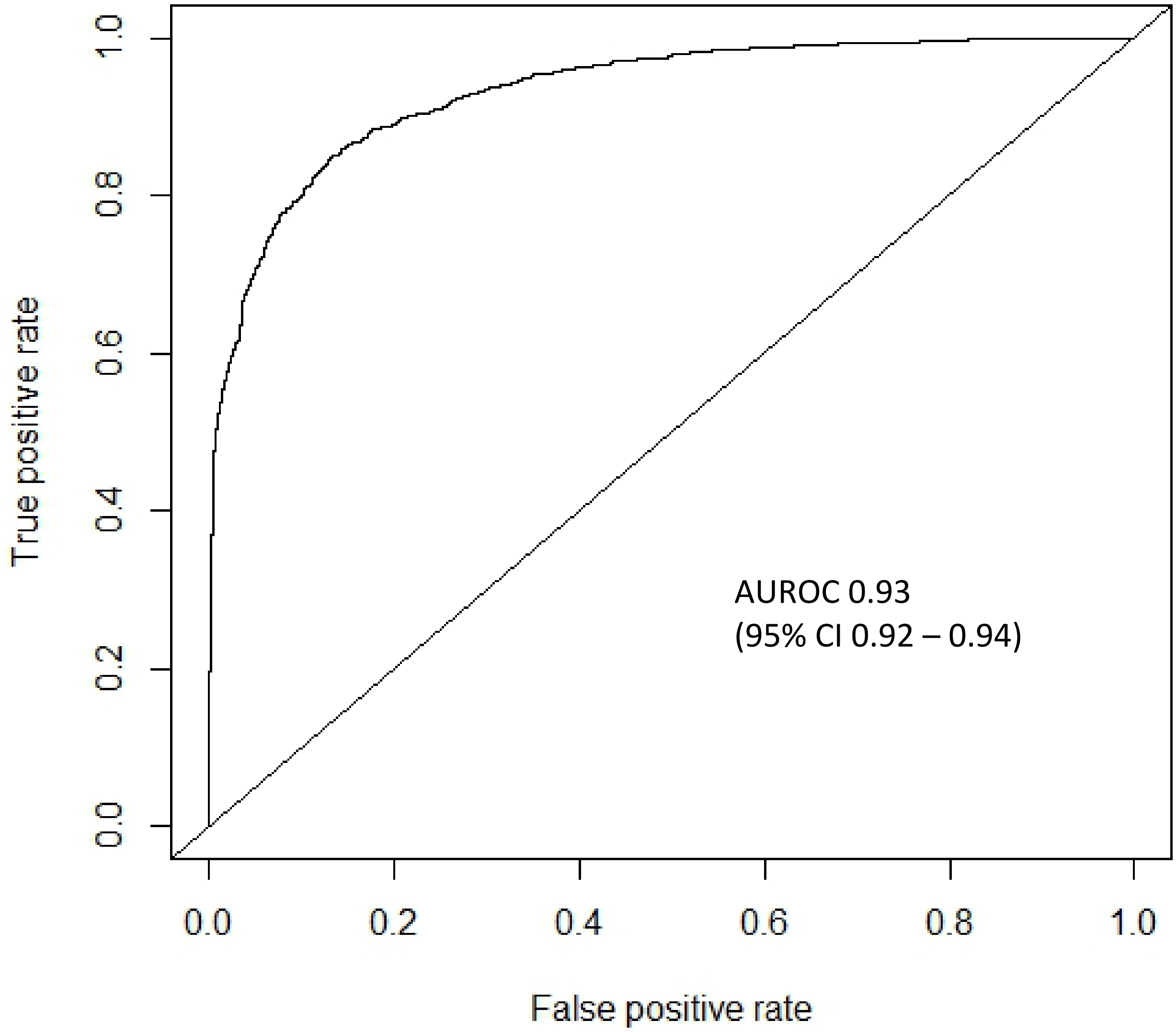


Figure 1. Receiver operating characteristic curve for predicting in-hospital major adverse cardiovascular events following open abdominal aortic aneurysm repair using Extreme Gradient Boosting (XGBoost) model. AUROC (area under the receiver operating characteristic curve), CI (confidence interval).

15 January 2025

Dear Research Committee,

Please find enclosed our manuscript “Using machine learning to predict outcomes following open abdominal aortic aneurysm repair.” Using data from a large clinical registry, we developed the first machine learning (ML) models to predict outcomes following open abdominal aortic aneurysm (AAA) repair.

In this paper, we demonstrate that our ML models are robust, explainable, and accurately predict in-hospital and 1-year outcomes following open AAA repair with better performance than existing tools and traditional statistical methods such as logistic regression. Therefore, our models have potential for important utility in the care of patients being considered for open AAA repair by guiding risk-mitigation strategies to improve outcomes.

The first author is Ben Li (Vascular Surgery Resident and SSTP trainee pursing a PhD at the Institute of Medical Science, University of Toronto). The supervisor is Dr. Mohammed Al-Omran (Vascular Surgeon, St. Michael’s Hospital, and Professor, University of Toronto). This work was completed in 2023 with ongoing validation in 2024-25 at St. Michael’s Hospital, Unity Health Toronto, University of Toronto.

The following are the authors’ contributions: concept and design (all authors), acquisition, analysis, or interpretation of data (all authors), drafting of the manuscript (Li), critical revision of the manuscript for important intellectual content (all authors), statistical analysis (Li), administrative, technical, or material support (Beaton, Eisenberg), and supervision (Al-Omran).

Thank you for your consideration.

Sincerely,

Ben Li, MD
Badr Aljabri, MD
Raj Verma, MD(c)
Derek Beaton, PhD
Naomi Eisenberg, PT, MEd, CCRP
Douglas S. Lee, MD, PhD
Duminda N. Wijeysundera, MD, PhD
Thomas L. Forbes, MD
Ori D. Rotstein, MD, MSc
Charles de Mestral, MD, PhD
Muhammad Mamdani, MPH, MA, PharmD
Graham Roche-Nagle, MD, MBA
Mohammed Al-Omran, MD, MSc

USING ARTIFICIAL INTELLIGENCE TO OPTIMIZE TRAUMA PATIENT TRANSFER AND RESOURCE UTILIZATION DURING MASS-CASUALTY INCIDENTS: DEVELOPMENT AND VALIDATION OF A NOVEL PLATFORM FOR SIMULATION TRAINING

Zhaoxun "Lorenz" Liu[1], Jay Han[2], Barbara Haas[3, 5], Matthew Guttman[3, 5], Kameela Alibhai[5], Homer Tien[3, 5], Jordyn Hurly[4], Joseph Sakran[4], Amin Madani[1, 5]

[1] Surgical Artificial Intelligence Research Academy, University Health Network, Toronto, Canada
[2] Department of Anesthesia and Pain Management, University Health Network, Toronto, Canada
[3] Department of Surgery, Sunnybrook Health Sciences Centre, Toronto, Canada
[4] Department of Surgery, Johns Hopkins University, Baltimore, USA
[5] Department of Surgery, University of Toronto, Toronto, Canada

INTRODUCTION

Mass casualty incidents (MCI) are disasters (e.g. natural disasters, explosions, chemical spills, plane crashes, terrorist attacks, military conflict) that overwhelm the local healthcare system and management agencies. MCIs require prompt assessment, triaging and transfer of patients to the hospital that is best equipped to accommodate the myriads of potential injuries. Decision-making by MCI commanders can be particularly challenging as they attempt to make critical decisions in a timely fashion while attempting to coordinate with the multiple team members at the disaster site and destination hospitals. This requires detailed understanding of patient factors (e.g. number of victims, mechanisms/types of injuries), hospital factors (e.g. distance from MCI site, available ICU beds/operating rooms/surgeons/ventilators) and transportation factors (e.g. available ambulances, helicopters) to optimize transfer decisions and patient outcomes. Given the relative rarity and uniqueness of each MCI, training healthcare workers in MCI response decision-making is limited mostly to tabletop simulation exercises. There is therefore a need for innovative methodologies for improving decision-making for MCI in a cost-effective and scalable manner. One potential solution is through the use of digital solutions and intelligent systems for providing end-users with simulation-based training and decision-support.

We aimed to 1) develop a novel digital platform to simulate MCI events, 2) train and validate an artificial intelligence (AI) algorithm that provides decision-support to accelerate and optimize patient transfer decision, and 3) determine whether it improves decisions amongst both trauma experts and non-experts in a simulated environment.

METHODS

Simulation Platform

MasTER (Mass-Casualty Trauma and Emergency Response) is an intelligent human-in-the-loop command dashboard accessible as a web application that simulates MCIs and provides end-users with a virtual environment to make transfer decisions (Figure 1). On the platform, users can assess the various trauma patients, injury mechanism, injury severity classification according to accepted standards for MCI, as well as the potential hospitals they can transfer to, including travel time and the various resources at each site (e.g. availability of intensive care unit beds, operating rooms, ventilators, blood products, etc.). The platform was designed to be capable of simulating MCIs in any geographic location, any combination of total number of trauma victims, injury mechanism/severity, available institutions and resources at each site. To optimize the fidelity of the simulation environment, the identification of trauma victims, and availability of transportation vehicles and hospital resources follows a sigmoid relationship as time progresses. Users can assign patients to be transferred to specific hospitals either through direct drag-and-drop function or by requesting AI-generated suggestions, which can be accepted or declined.

Machine Learning Model

To facilitate patient transfer decisions, an AI algorithm we trained using deep reinforcement learning (DRL). DRL [1] is a machine learning paradigm that enables continuous learning through trial-and-error interactions with complex simulated environments while optimizing long-term outcomes. This approach proves advantageous when structured training data are limited and has shown utility in healthcare setting [2,6,7,8]. Deep Learning (DL) [3] employs multi-layered artificial neural networks to extract high-dimensional features and generate predictions. DRL [4], on the other hand, combines the strengths of both, presenting an effective solution through its capacity for ongoing adaptation in complicated scenarios. DRL's capability to manage high-dimensional state and action spaces aligns with the multifaceted nature of MCIs, while simulation addresses the scarcity of organized historical incident information. For this model, we chose a Proximal

Policy Optimization (PPO)-based DRL approach [5] for MasTER's algorithm. The AI agent was trained on extensive simulated MCI scenarios (n=10,000), representing diverse casualty volumes (ranging from 10-500 patients), injury patterns, and regional hospital resource configurations. The modelling of simulated MCIs strictly follows the existing rapid trauma triage protocols, such as color codes for patient severity and levels for hospital capability [14]. At its core, the agent optimized a multi-objective reward function prioritizing survival probability while considering transport duration, facility capacity constraints, and specialty care requirements. Once the model was developed, it was integrated into the simulation platform as an add-on feature where end-users can summon the model to provide a recommendation on the optimal destination hospital for any given patient.

Validation

To evaluate whether MasTER facilitates MCI management, we conducted user studies using two distinct simulation exercises: a Standard level (20 patients) and Complex level (60 patients), both in the Greater Toronto Area. Users participated in both simulation in two iterations: 1) Human-only (no AI available to assist) and 2) Human+AI (human-in-the-loop approach where AI assistance was available at their discretion). In the Human+AI setting, participants could request AI-generated suggestions for patient assignments which they could either accept or decline. We also tested the AI model as a standalone fully autonomous agent making decisions (AI-only). Participants (including trauma surgeons and non-trauma surgeons) were recruited to complete the Human-only and Human+AI simulations. Prior to starting the simulation exercises, all participants underwent a training module to gain familiarity with the platform.

Outcomes

Our evaluation framework incorporated both quantitative and qualitative measures to provide a comprehensive assessment of the MasTER platform. The system automatically collected quantitative metrics including total completion time and patient survival rates. Survival and mortality were determined according to preestablished benchmarks for given injury severity requiring transfer to specific hospitals with the necessary resources within a given time frame. For the Human+AI condition, we tracked the acceptance rate of AI suggestions to understand user trust and system reliance. For qualitative assessment, we measured workload using the NASA Task Load Index (NASA-TLX) [9] and system usability via the System Usability Scale (SUS) [10]. We also assessed user perception on the utility and value of the tool. These post-condition evaluations captured immediate impressions and experiences while minimizing recall bias. All questionnaires were administered and collected via REDCap [11, 12].

RESULTS

There was a total of 30 participants, including 6 expert trauma surgeons from a high-volume Level 1 trauma hospital, and 24 non-experts. For the entire cohort, there were significant differences across conditions (Human-only, Human+AI, and AI-only) for all performance metrics. For completion time (Figure 3a), repeated measures ANOVA [13] showed a significant main effect ($F(2,87) = 892.31$, $p < .001$, $\eta^2p = .943$). Post-hoc Tukey's HSD tests indicated significant differences between all pairs of conditions ($p < .001$). The Human+AI condition demonstrated substantially improved performance compared to the Human-only condition ($d = 4.72$, 95% CI [4.23, 5.21]). There was a significant improvement in simulated mortality rates (Figure 3b), using AI assistance for the Standard level ($t(29) = 7.82$, $p < .001$; Human: M = 1.00%, SD = 2.03%; Human+AI: M = 0.17%, SD = 0.91%). This improvement was even more pronounced during Complex scenarios where there were more patients to be traiged ($t(29) = 11.23$, $p < .001$; Human: M = 6.37%, SD = 1.71%; Human+AI: M = 2.83%, SD = 1.02%). Match rates (Figure 3c) similarly improved with AI assistance, with the Standard level achieving near-perfect scores in the Human+AI condition (M = 98.13%, SD = 0.31%) compared to Human-only (M = 90.07%, SD = 2.76%; t(29) = 15.82, p < .001). Error rates (Figure 3d) showed consistent improvement across both difficulty levels, with the Human+AI condition reducing errors by 42.3% in the Standard level (t(29) = 16.34, p < .001) and 71.2% in the Complex level (t(29) = 19.87, p < .001) compared to Human-only. The AI-only condition achieved near-perfect scores across all metrics, establishing a theoretical performance ceiling.

When analyzing expert versus non-expert performance, we found that non-experts were unable to match experts' performance in the Human-only condition across all metrics. However, when provided with AI assistance (the Human+AI condition), non-experts demonstrated remarkable improvement, achieving metrics that surpassed experts' Human-only performance.

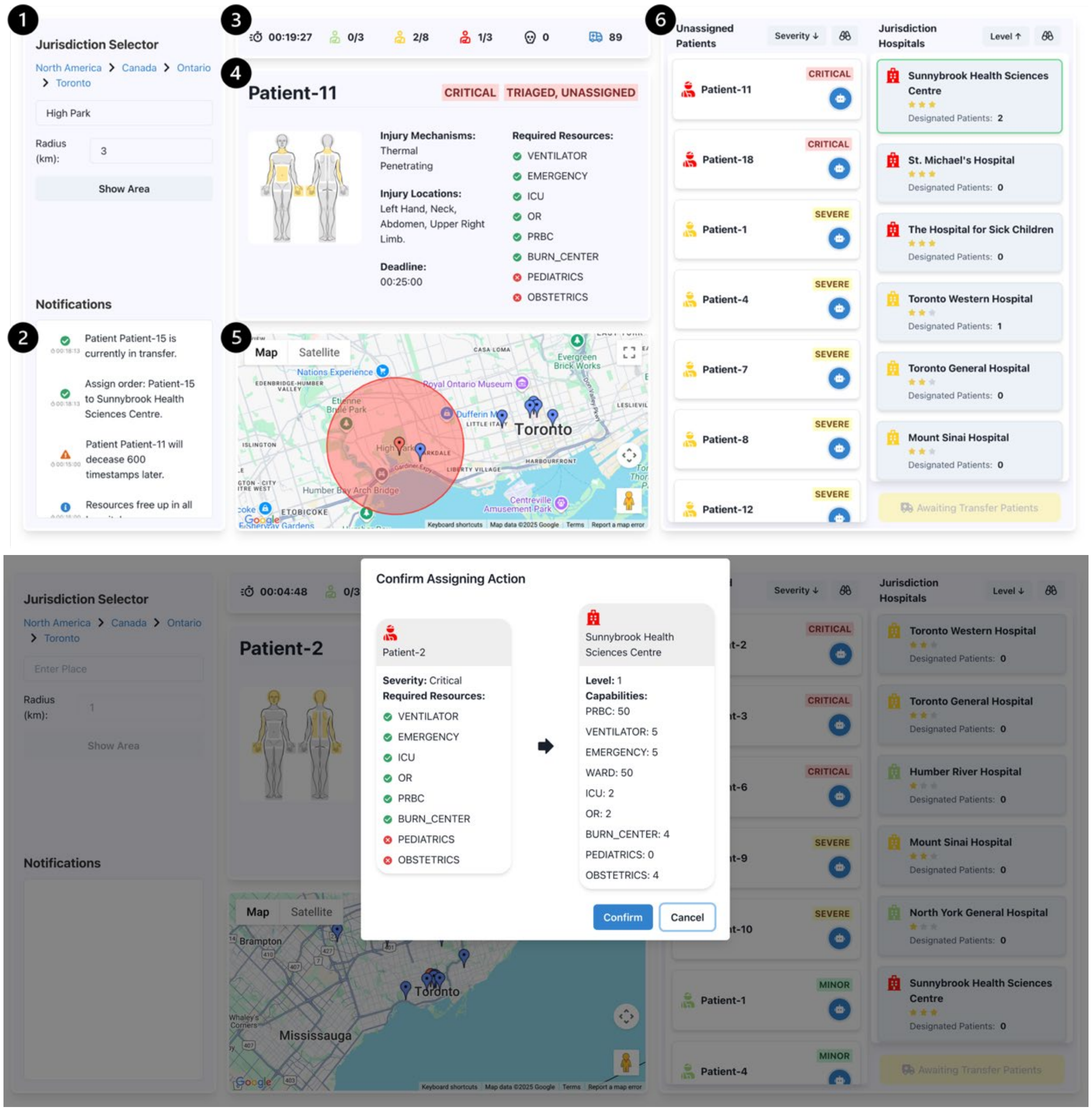
Qualitative analysis of NASA-TLX scores indicated significantly lower perceived workload in the Human+AI condition (M = 31.4, SD = 3.85) compared to Human-only (M = 63.7, SD = 7.2; t(29) = 17.92, p < .001). The System Usability Scale (SUS) score for the Human+AI system was exceptional at 87.87 (SD = 2.54), placing it in the 95th percentile of evaluated systems.

Discussion

In simulated MCI, the use of the AI model significantly improved triage decisions, with substantial mortality rate reductions (83% for the Standard level, 55.6% for the Complex level; p < .001) and error rate reductions (42.3-71.2%). The large effect size for completion time (d = 4.72) shows both statistical and practical significance. The DRL model performed even better autonomously than with human intervention. Furthermore, MasTER's features demonstrated strong utility, with NASA-TLX scores showing 50.7% reduced workload (p < .001) and exceptional usability (SUS: 87.87, 95th percentile). Features provided crucial support in complex scenarios, particularly in the Complex level where there are many victims and decisions become extremely challenging (error reduction: p < .001). Finally, trauma experts generally found MasTER useful, robust, and efficient, as evidenced by high satisfaction scores and significant time improvements (25.49% faster for the Standard level, 45.35% for the Complex level; p < .001). Low error rates in Human+AI condition (71.2% reduction in the Complex level) indicate that this system could be effectively integrated into existing workflows.

CONCLUSIONS

We developed a novel web application simulation platform with AI decision-support (MasTER) and showed that it significantly improves decision-making metrics within a simulated environment while reducing cognitive burden during MCI management. Moreover, our study demonstrates that increasing AI involvement produces better, more stable, and predictable transfer decisions. While MasTER could successfully address current MCI management challenges, future development should incorporate more comprehensive resources (e.g., neurosurgery availability, helicopter transportation option), balance workload between same-level hospitals, automation of input for patient profiles and hospital resources, and conduct longitudinal studies to assess patient and system outcomes.



High Park

Radius (km): 3

Patient Patient-15 is currently in transfer. (00:18:13)

Assign order: Patient-15 to Sunnybrook Health Sciences Centre. (00:18:13)

Patient Patient-11 will decease 600 timestamps later. (00:15:00)

Resources free up in all

Patient-11 CRITICAL TRIAGED, UNASSIGNED

Injury Mechanisms:
Thermal
Penetrating

Injury Locations:
Left Hand, Neck, Abdomen, Upper Right Limb.

Deadline:
00:25:00

Required Resources:
- ✓ VENTILATOR
- ✓ EMERGENCY
- ✓ ICU
- ✓ OR
- ✓ PRBC
- ✓ BURN_CENTER
- ☒ PEDIATRICS
- ☒ OBSTETRICS

Unassigned Patients	
Patient-11	CRITICAL
Patient-18	CRITICAL
Patient-1	SEVERE
Patient-4	SEVERE
Patient-7	SEVERE
Patient-8	SEVERE
Patient-12	SEVERE

Jurisdiction Hospitals	Designated Patients
Sunnybrook Health Sciences Centre ★★★	2
St. Michael's Hospital ★★★	0
The Hospital for Sick Children ★★★	0
Toronto Western Hospital ★★☆	1
Toronto General Hospital ★★☆	0
Mount Sinai Hospital ★★☆	0

Jurisdiction Selector

Enter Place

Radius (km): 1

Patient-2

Confirm Assigning Action

Patient-2

Severity: Critical
Required Resources:
- ✓ VENTILATOR
- ✓ EMERGENCY
- ✓ ICU
- ✓ OR
- ✓ PRBC
- ✓ BURN_CENTER
- ☒ PEDIATRICS
- ☒ OBSTETRICS

➡

Sunnybrook Health Sciences Centre

Level: 1
Capabilities:
PRBC: 50
VENTILATOR: 5
EMERGENCY: 5
WARD: 50
ICU: 2
OR: 2
BURN_CENTER: 4
PEDIATRICS: 0
OBSTETRICS: 4

Unassigned Patients	
…t-2	CRITICAL
…t-3	CRITICAL
…t-6	CRITICAL
…t-9	SEVERE
…t-10	SEVERE
Patient-1	MINOR
Patient-4	MINOR

Jurisdiction Hospitals	Designated Patients
Toronto Western Hospital ★★☆	0
Toronto General Hospital ★★☆	0
Humber River Hospital ★☆☆	0
Mount Sinai Hospital ★★☆	0
North York General Hospital ★☆☆	0
Sunnybrook Health Sciences Centre ★★★	0

Figure 1 (Top) The user interface of MasTER has 6 major components: ❶ Jurisdiction Selector enables hierarchical navigation of geographical responsibility areas with radius-based refinement; ❷ Notifications delivers color-coded, timestamped updates on system events for quick assessment; ❸ Status Bar provides real-time metrics including elapsed time, unassigned patients by severity, mortality count, and available ambulances; ❹ Detail Panel displays comprehensive information about selected patients (severity, injuries, resource needs) or hospitals (level, capabilities, distance); ❺ Interactive Map visualizes incident site and hospitals with multiple viewing options for improved spatial awareness.; ❻ Draggable Action Panel presents two synchronized lists: unassigned patients and hospitals in the selected jurisdiction by default. Patients are color-coded by severity (Critical in red, Severe in yellow, Minor in green, and Deceased in gray), while hospitals are differentiated by their trauma level (Level 1-3). Users can assign patients either through direct drag-and-drop interactions or by requesting AI-generated suggestions, which can be accepted or declined; (Bottom) an example of AI suggestion is given.

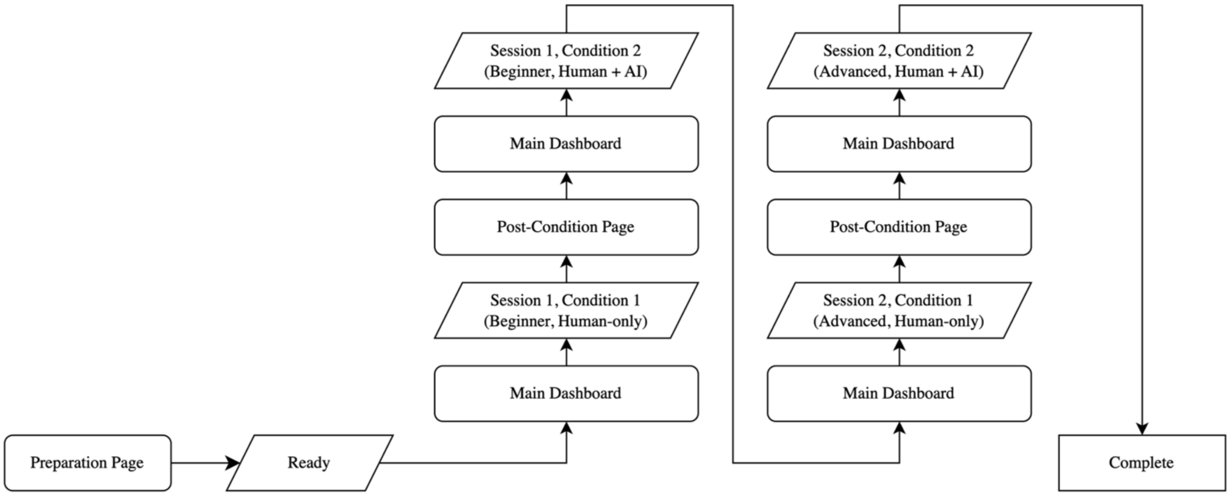
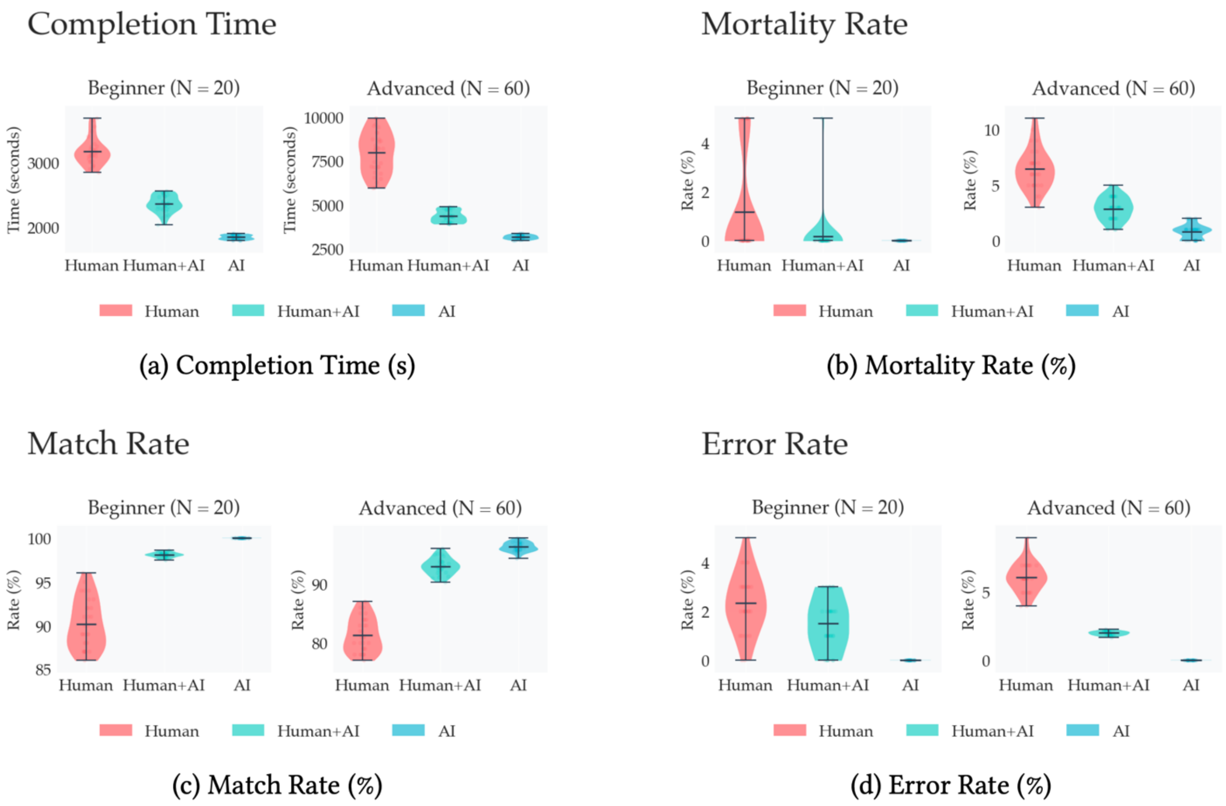


Figure 2 Procedure of the user study.



(a) Completion Time (s)

(b) Mortality Rate (%)

(c) Match Rate (%)

(d) Error Rate (%)

Figure 3 Quantitative results from the user study.

REFERENCES

1. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
2. Jayaraman, P., Desman, J., Sabounchi, M. et al. (2024). A Primer on Reinforcement Learning in Medicine for Clinicians. npj Digit. Med. 7, 337.
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
4. Li, Y. (2017). Deep Reinforcement Learning: An Overview. ArXiv, abs/1701.07274.
5. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint1 arXiv:1707.06347.
6. Coronato, A., Naeem, M., De Pietro, G., & Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. Artificial Intelligence in Medicine, 109, 101964.
7. Ji, S., Zheng, Y., Wang, Z., & Li, T. (2019). A deep reinforcement learning-enabled dynamic redeployment system for mobile ambulances. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(1), Article 15.
8. Liu, K., Li, X., Zou, C. C., Huang, H., & Fu, Y. (2020). Ambulance dispatch via deep reinforcement learning. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems* (SIGSPATIAL '20) (pp. 123–126). Association for Computing Machinery. https://doi.org/10.1145/3397536.3422204
9. Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Human mental workload (pp. 139–183). North-Holland.
10. Brooke, J. (1986) SUS—A Quick and Dirty Usability Scale. Usability Evaluation in Industry, 189-194.
11. Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J. G. (2009). Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics, 42(2), 377–381.
12. Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., Duda, S. N., & REDCap Consortium. (2019). The REDCap consortium: Building an international community of software partners. Journal of Biomedical Informatics. Advance online publication.
13. Girden, E. R. (1992). ANOVA: Repeated measures. Sage Publications, Inc.
14. Clarkson, L., & Williams, M. (2023). EMS Mass Casualty Triage. In StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK459369/

DEEP BRAIN STIMULATION OF THE NUCLEUS ACCUMBENS FOR SEVERE SELF-INJURIOUS BEHAVIOURS IN CHILDREN: A PHASE I PILOT TRIAL

Karim Mithani (SSTP)[1,2], Carolina Gorodetsky[3], Sara Breitbart[1], Han Yan[4], Kristina Zhang[5], Flavia Venetucci Gouveia[6], Nebras Warsi[1,2], Hrishikesh Suresh[1,2], Simeon M. Wong[2], Joelene Huber[7], Elizabeth N. Kerr[8], Abhaya V. Kulkarni[1,9], Margot J. Taylor[6,8,10], Louis Hagopian[11], Alfonso Fasano[3,5,12,13], George M. Ibrahim[1,2,5,6]

[1] Division of Neurosurgery, The Hospital for Sick Children, Department of Surgery, University of Toronto, Toronto, Ontario, Canada
[2] Institute of Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada
[3] Division of Neurology, The Hospital for Sick Children, Toronto, Ontario, Canada
[4] Department of Family and Community Medicine, University of Toronto, Toronto, Canada
[5] Institute of Medical Science, University of Toronto, Toronto, Canada
[6] Program in Neuroscience and Mental Health, Hospital for Sick Children, Toronto, Canada
[7] Division of Pediatric Medicine and Developmental Pediatrics, Department of Pediatrics, Hospital for Sick Children, University of Toronto, Toronto, Canada
[8] Department of Psychology, The Hospital for Sick Children, Toronto, Canada
[9] Institute of Health Policy, Management and Evaluation, University of Toronto
[10] Department of Diagnostic Imaging, Hospital for Sick Children, Toronto, Canada
[11] Neurobehavioural Unit, Department of Behavioural Psychology, Kennedy Krieger Institute, Baltimore, USA
[12] Edmond J. Safra Program in Parkinson's Disease, Morton and Gloria Shulman Movement Disorders Clinic, Toronto Western Hospital, UHN, Toronto, Canada
[13] Krembil Brain Institute, Toronto, Canada

The authors have decided not to make the research results available at this time and will provide updates as soon as the results can be shared.

DEVELOPMENT AND PERFORMANCE OF VASC.AI: A VASCULAR SURGERY AI PLATFORM FOR CLINICAL DECISION SUPPORT

Arshia Javidan[1], Tiam Feridooni[1], Paul Akouris[2], Kaan Balta[3], Allen Li[1], Daniyal Mahmood[2], Andrew Dueck[1], Mark Wheatcroft[1], David Szalay[1]
[1]Division of Vascular Surgery, Department of Surgery, University of Toronto, Toronto, Ontario
[2]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario
[3]Schulich School of Medicine, University of Western Ontario, London, Ontario

INTRODUCTION

The rapid evolution of artificial intelligence has transformed numerous facets of medicine, yet its application in highly specialized fields such as vascular surgery remains limited. General-purpose large language models (LLMs) like ChatGPT have demonstrated significant potential in processing vast quantities of text and providing clinically relevant information. However, these models are plagued by challenges—most notably "hallucinations" and an inherent lack of domain specificity—which compromise their reliability in complex clinical decision-making. In vascular surgery, where precision and adherence to updated guidelines are critical, even minor inaccuracies can have significant clinical consequences.

To address these challenges, we developed VASC.AI, a dedicated AI platform that harnesses retrieval-augmented generation (RAG) techniques. By integrating a curated database of over 250,000 vascular surgery abstracts, clinical guidelines, and landmark trial results, VASC.AI is engineered to deliver evidence-based and contextually robust recommendations. This study comprehensively evaluates VASC.AI's performance relative to baseline ChatGPT models across four distinct domains: (1) guideline-based questions, (2) multiple-choice questions from the Vascular Education and Self-Assessment Program (VESAP), (3) certification-level exam questions in peripheral arterial disease (PAD) and cerebrovascular disease, and (4) complex clinical scenarios in cerebrovascular disease. Our aim is to demonstrate that by augmenting a LLM with vascular surgery-specific literature, VASC.AI not only mitigates the pitfalls of generalized LLMs but also substantially enhances clinical decision support.

METHODS

Development of VASC.AI

VASC.AI was developed using a retrieval-augmented generation (RAG) framework that integrates a specialized, domain-specific knowledge base with a state-of-the-art large language model. This knowledge base comprises over 250,000 curated vascular surgery abstracts, clinical guidelines, and landmark trial publications, all of which were vectorized and embedded using advanced natural language processing techniques. This integration enables efficient retrieval of contextually relevant, evidence-based information, allowing VASC.AI to generate precise, context-aware responses tailored to the complexities of vascular surgery (Figure 1).

Study Design and Data Collection

We implemented a multi-phase evaluation protocol comparing VASC.AI with pre-existing large language models, including ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Gemini, and Copilot. Varying difficulties and complexities of questions were inputted into each model, the responses to which were assessed by two blinded vascular surgery experts. The evaluation was structured around four question sets.

Guideline-Based Questions

A series of 40 questions was generated directly from current vascular surgery guidelines relating to four domain areas of vascular surgery (carotid artery stenosis, visceral artery aneurysms, chronic limb-threatening ischemia, and abdominal aortic aneurysms). These questions were administered to each (ChatGPT-3.5, ChatGPT-4, and VASC.AI) and responses were evaluated for comprehensiveness and accuracy according to a validated Likert scale. Scores of greater than a 3 on the 5-point Likert scale were graded as appropriate.

VESAP Multiple-Choice Questions

We used 244 validated multiple-choice questions drawn from six modules of the VESAP-5 curriculum, covering areas including aortoiliac, cerebrovascular, lower extremity, renal/mesenteric, vascular medicine, and venous disease. Questions were input into each model (ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, and VASC.AI). Incorrect responses were categorized as either logical errors or information errors.

Certification-Level Exam Questions

Free-text questions simulating certification exam conditions were developed by expert panels, focusing on clinical scenarios in PAD and cerebrovascular disease. These were inputted into ChatGPT3.5 and ChatGPT4 and the VASC.AI augmented version of each model, where they were assessed using a validated 5-point Likert scale that measured diagnostic accuracy, comprehensiveness, and clarity.

Complex Clinical Scenarios

A set of realistic clinical cases was developed to simulate complex cases encountered in vascular surgery, specifically involving complex cerebrovascular diseases. Each case required the formulation of a differential diagnosis, a comprehensive workup plan, and a detailed treatment recommendation. These cases were input into the most powerful publicly available large language models at the time, ChatGPT4, Gemini, and Copilot, as well as VASC.AI. Evaluations were based on expert scoring of diagnostic accuracy, thoroughness of the workup, clarity in outlining optimization strategies, and relevance of the treatment options.

Statistical Analysis

All statistical analyses were performed using unpaired t-tests or one-way analysis of variance (ANOVA) followed by post hoc pairwise t-tests, with significance defined as p < 0.05. Performance data are reported as means with standard deviations.

RESULTS

Guideline-Based Questions

The baseline models showed a progression in performance when answering guideline-derived questions. ChatGPT-3.5 achieved an accuracy of 32.5% (13/40 correct recommendations) and ChatGPT-4 achieved an accuracy of 95% (38/40 correct recommendations, P < 0.001). In comparison, VASC.AI achieved an accuracy of 100%, providing an appropriate recommendation to every guideline-based question that was asked.

VESAP Multiple-Choice Questions

In the VESAP evaluation, ChatGPT-3.5, ChatGPT-4, and ChatGPT-4o demonstrated correct response rates of 55.3% (SD 4.3%), 69.0% (SD 4.9%), and 77.7% (SD 7.6%) respectively. VASC.AI outperformed these models with a correct response rate of 93.8% (SD 2.4%, p < 0.001), indicating its enhanced ability to interpret and apply current vascular standards of care to VESAP questions.

Certification-Level Exam Questions

Baseline models provided modest responses to certification-level free-text questions. ChatGPT-3.5 had a mean Likert score of 2.3 (SD 0.8) on PAD-related questions, while ChatGPT-4 scored 3.3 (SD 0.5). When augmented with vascular-specific content, the performance of these models improved significantly: the VASC.AI-enhanced ChatGPT-3.5 achieved a mean score of 3.9 (SD 0.6), and VASC.AI-enhanced ChatGPT-4 reached 4.8 (SD 0.4). Similar statistically significant improvements were observed in responses to cerebrovascular exam questions, with p-values of less than 0.0001 in pairwise comparisons (Figure 2).

Complex Clinical Scenarios

VASC.AI achieved the highest Overall Usefulness in Decision-Making score (4.93 ±0.26), significantly outperforming ChatGPT-4o (3.80 ± 1.26, *p* = 0.026), Gemini (3.07 ± 1.22, *p* = 0.0008) and Copilot (3.13 ± 1.19, *p* = 0.00014, Figure I). Regarding Relevance of Treatment Options, VASC.AI also had the highest scores in this category (4.80 ± 0.41), outperforming ChatGPT-4o (3.60 ± 1.45, *p* = 0.035), Gemini (3.07 ± 1.33, *p* = 0.00097), and Copilot (3.07 ±1.22, *p* = 0.00097). For Clarity in Medical Optimization

Plans, there was no statistically significant difference between VASC.AI (4.73 ± 0.46) and ChatGPT-4o (4.60 ± 0.63, *p* = 0.98), but VASC.AI did outperform both Gemini and Copilot (both 3.80 ± 1.21, p = 0.042). VASC.AI outperformed Gemini in the Thoroughness of Workup category (4.87 ± 0.35 vs. 4.07 ± 0.96, *p* = 0.014), with no other statistically significant differences in this category amongst the language models. There were no statistically significant differences in the Accuracy of the Differential Diagnosis score amongst the language models.

CONCLUSIONS

VASC.AI represents a significant advancement in the use of artificial intelligence within vascular surgery. By integrating a comprehensive, domain-specific knowledge base with advanced retrieval-augmented generation techniques, VASC.AI achieves complete accuracy on guideline-based questions and substantially outperforms general-purpose LLMs on multiple-choice assessments, certification-level free-text questions, and complex clinical scenarios. The data indicate that VASC.AI can deliver precise, evidence-based recommendations that support enhanced clinical decision-making and improve educational outcomes in vascular surgery.

Moreover, the rapid pace of developments in artificial intelligence underscores the critical importance of adapting these technologies to meet specialized clinical needs. Our study serves as a robust proof of concept that retrieval-augmented generation not only improves the performance of large language models but also deepens their domain expertise. As AI technology continues to evolve, the integration of specialized knowledge bases—as exemplified by VASC.AI—will be vital in ensuring these tools contribute effectively to improved patient care, streamlined clinical workflows, and optimized education and training. Future steps will focus on incorporating multimodal data inputs, such as imaging, refining logical reasoning capabilities, and expanding the system's applicability to cover a broader spectrum of vascular conditions. These advancements will help drive further improvements in clinical efficiency and decision-making, ultimately paving the way for a new era of domain-specific AI in vascular surgery.

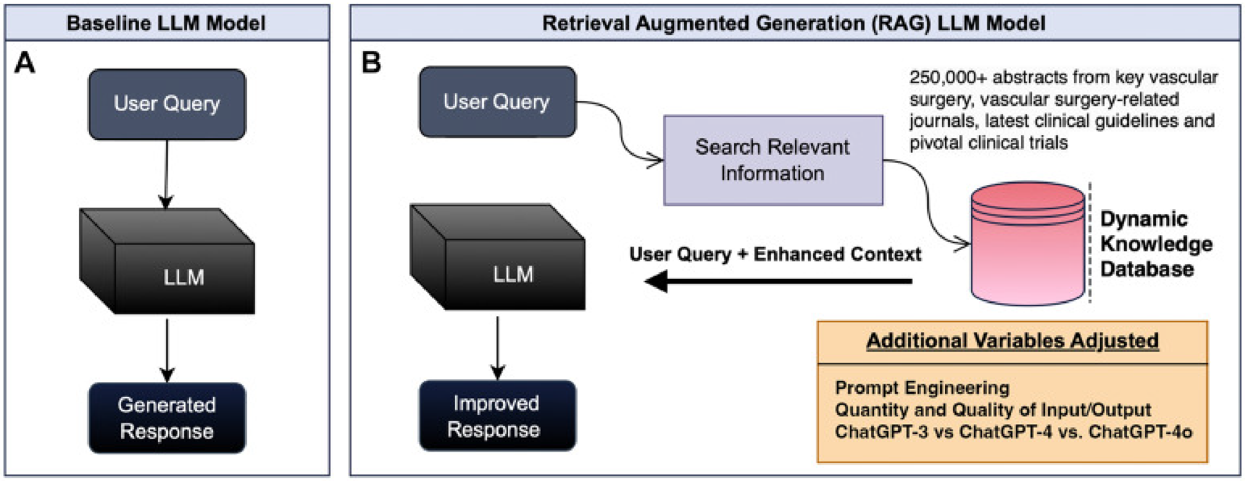

Baseline LLM Model
A
User Query
LLM
Generated Response
Retrieval Augmented Generation (RAG) LLM Model
B
User Query
Search Relevant Information
250,000+ abstracts from key vascular surgery, vascular surgery-related journals, latest clinical guidelines and pivotal clinical trials
Dynamic Knowledge Database
User Query + Enhanced Context
LLM
Improved Response
Additional Variables Adjusted
Prompt Engineering
Quantity and Quality of Input/Output
ChatGPT-3 vs ChatGPT-4 vs. ChatGPT-4o

Figure 1. A comparison of the functionality of large language models (LLM) in its base form without retrieval augmented generation (RAG) (A), and VASC.AI (B). In panel A, a user will generate a query which is inputted into the base LLM, which generates a response without additional resources or context. In panel B, the same query is matched against a comprehensive database of key vascular surgery resources, where additional context-appropriate information is gathered and fed back to the same LLM which now delivers a response improved with RAG.

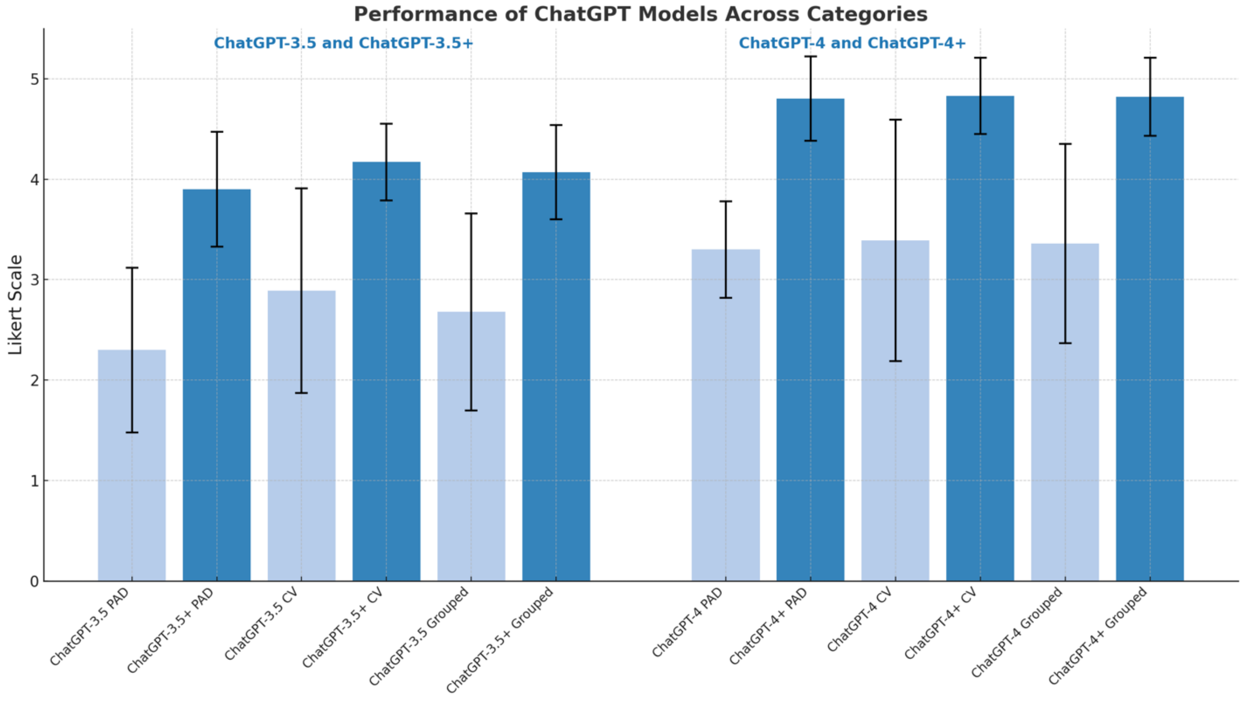


Figure 2. Performance comparison of baseline ChatGPT3.5 and ChatGPT4 models with their VASC.AI-enhanced counterparts (ChatGPT3.5+ and ChatGPT4+). Scores were evaluated using a 5-point Likert scale for accuracy and comprehensiveness on vascular surgery exam-level questions covering peripheral arterial and cerebrovascular diseases. Scores are additionally grouped to represent performance on both peripheral arterial and cerebrovascular disease.

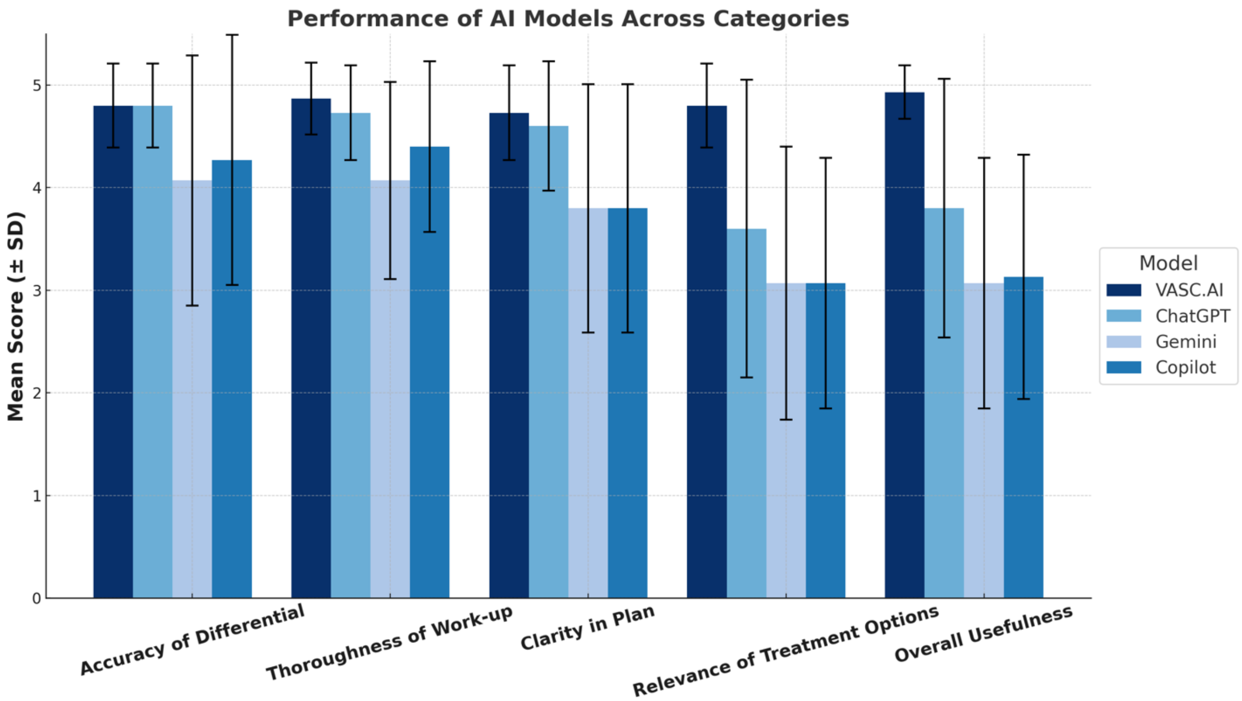


Figure 3. Performance of AI models across five evaluation categories. The mean scores (± standard deviation) for VASC.AI, ChatGPT-4o, Gemini, and Copilot are shown in the categories of Accuracy of Differential, Thoroughness of Work-up, Clarity in Plan, Relevance of Treatment Options, and Overall Usefulness in Decision-Making. VASC.AI demonstrates consistently higher performance, particularly in Overall Usefulness and Relevance of Treatment Options, compared to general-purpose models, highlighting its value as a domain-specific AI platform.