INTEGRATING CHAT GPT IN ANESTHESIA

Implementation of Large Language Models in Anesthesia to Answer Patient's Questions During Pre-Anesthesia Visits: A Prospective, Observational Study

CLINICAL RESEARCH PROTOCOL v.1.0:03-Oct-2024

Qualified/Principal Investigator:

Arnaud Romeo, Mbadjeu Hondjeu MD, MSc Candidate
Department of Anesthesiology & Pain Medicine
The Ottawa Hospital Research Institute
725 Parkdale Avenue
Ottawa, ON K1Y 4E9

Co-Principal Investigators:

Daniel McIsaac, MD, MPH, FRCPC

Sponsor: Ottawa Hospital Research Institute

Funders: University of Ottawa Department of Anesthesiology & Pain Medicine (uODAPM)

Coordinating Center:

Ottawa Hospital Research Institute
Department of Anesthesiology & Pain Medicine Research
1053 Carling Avenue
Room D106b
Ottawa, ON K1Y 4E9

Approval:

______________________________ ______________
Signature of Nominated Qualified/Principal Investigator **Date (DD-MMM-YYYY)**

DOCUMENT HISTORY

Version	Version Date	Changes

PROTOCOL SYNOPSIS

Study Title:	Implementation of Large Language Models in Anesthesia to Answer Patient’s Questions During Pre-Anesthesia Visits: A Prospective, Observational Study
Protocol Short Title:	Integrating ChatGPT in Anesthesia
Protocol Version:	1.0
Protocol Date:	03-Oct-2024
Qualified/Principal Investigator:	**Arnaud Romeo, Mbadjeu Hondjeu MD, MSc Candidate** Department of Anesthesiology & Pain Medicine The Ottawa Hospital, General Campus 501 Smyth Road Ottawa, ON K1H 8L6
Co-Principal Investigators	Daniel McIsaac, MD, MPH, FRCPC
Sponsor:	Ottawa Hospital Research Institute (OHRI)
Funders:	University of Ottawa Department of Anesthesiology & Pain Medicine (uODAPM)
Coordinating Center:	Ottawa Hospital Research Institute Department of Anesthesiology & Pain Medicine Research 1053 Carling Avenue Room D106b Ottawa, ON K1Y 4E9

Study Objectives:	• Estimate ChatGPT non-inferiority to clinician's responses to patient queries by examining gain in knowledge, actionable information, completeness, patient satisfaction, and perceived empathy in the clinical setting. • Estimate the incidence of hallucinations (plausible sounding but incorrect or nonsensical answers) by ChatGPT in responding to clinical queries. • Compare the performance of ChatGPT across population sub-groups (English and French, different sex and gender, age sub-groups as defined by WHO, and education level as defined by UNESCO). • Engage and inform patients and key health system users via Integrated knowledge translation (iKT) throughout the research process.
Study Design:	A prospective, observational cross-sectional study using a non-inferiority design
Number of Participants:	190
Study Population:	Adults having an in-person preoperative anesthesiology consultation before elective surgery at The Ottawa Hospital
Inclusion Criteria:	1. Age ≥ 18 years 2. Elective non-cardiac surgery 3. In-person preoperative anesthesiology consultation 4. Ability to participate and provide informed consent independently.
Exclusion Criteria:	1. Unable to communicate in English or French 2. Nurse consultation only
Follow-Up Duration:	All follow-ups will be completed during the pre-admission unit appointment

1. BACKGROUND AND RATIONALE

1.1 Shortage of anesthesiologists in Canada

Globally, nationally, and provincially, there is a persistent shortage of anesthesiologists.[1] This shortage is a key barrier to addressing surgical wait times, especially as we emerge from the health system shocks of the COVID-19 pandemic. In addition to staffing each operating room, effective preoperative assessment by anesthesiologists is crucial to address patients' informational needs, decrease day of surgery cancellations, and reduce morbidity and mortality.[2] While different solutions have been proposed to tackle this issue, the crux and swift response to the problem lies in how we can optimize the allocation of scarce and highly trained anesthesiologists while maintaining an optimum level of care for patients. Artificial Intelligence (AI) programs like ChatGPT might help by answering common questions that patients have, which could allow anesthesiologists to spend their limited time addressing growing surgical waitlists and resolving complex and personal issues for each patient.

1.2 Promises and limitations of large language models.

Large language models (LLMs) are natural language processing computer programs capable of predicting and generating human language using artificial neural networks.[3] They work by taking input text and repeatedly predicting the next word based on pre-trained data. The models offer advanced data analysis and decision support in academic and industrial domains, with notable examples such as OpenAI's ChatGPT(Open AI 2022), Google's Pathways Language Model (PaLM 2) in Bard (Google 2023),[4] as well as the more recent multimodal Google's Gemini.[5]

ChatGPT (Generative Pre-trained Transformer)[6] is an LLM capable of capturing human language nuances and generating contextually relevant responses across prompts based on training data up to September 2021.[7] However, limitations, such as hallucinations (plausible-sounding but incorrect or nonsensical answers) and limited knowledge of events after September 2021, may limit accuracy in critical settings.[8] OpenAI introduced GPT-4 in March 2023, enhancing reliability, memory, multilingual capabilities, and steerability, reducing unapproved prompts and facts fabrication. GPT-4 can use Bing for up-to-date answers on the internet.[9] This LLM is a transformer-based model trained using a mixture of objectives with improved multilingual and reasoning capabilities.[10] Based on experience to date, the availability and performance of LLMs is expected to increase substantially over time.

1.3 Potential implementation of large language models in health care

Since inception, LLMs have enabled enterprises to design innovative products like automated customer service, report generation, and knowledge management.[11] ChatGPT has been reported to be capable of answering UK Royal College of Anaesthetists practice questions at a level close to the pass mark,[12] while also achieving the passing criteria for the US Medical Licensing Examination,[13] suggesting potentially impactful applications in the health care settings. A recent cross-sectional study found that ChatGPT responses to medical queries were largely judged as adequate by physicians[14] Unfortunately, this and related studies possess substantive limitations such as the absence of patient perspectives, queries, or ratings and the inability to assess hallucinations and refine or iterate ambiguous questions to provide the most accurate response. Robust, patient-partnered research will be required to address these gaps before considering the implementation LLMs in clinical settings.

1.4 Current Landscape of LLMs in the Perioperative Setting

The implementation of LLMs in perioperative settings remains limited. A single-blind, randomized controlled trial aimed to evaluate the effects of ChatGPT-3.5, on preoperative anxiety reduction and patient satisfaction in adults undergoing surgery demonstrated that ChatGPT intervention reduced preoperative anxiety compared with the control group; however, no overall difference in the Japanese State-Trait Anxiety Inventory scores were reported.[15] One study suggested that ChatGPT achieves comparable preoperative risk scoring to humans with less variability,[16] while a recent report demonstrated that ChatGPT provides quality and empathetic responses to patient questions posed in an online forum.[17] However, these comparisons and those described above are limited and are likely biased. Specifically, the human-generated responses that ChatGPT was compared to appear to have had different intended audiences. Furthermore, online forum questions cannot be compared to patients' clinical queries. Hence, the performance of LLMs in the clinical setting has yet to be robustly evaluated. Moreover, when considering LLMs in clinical practice, hallucination incidence is a crucial parameter that needs to be assessed while in a bilingual country like Canada, performance in different languages (especially French) are key parameters that have not yet been reported. The proposed prospective, single center observational cross-sectional study will produce findings that will directly inform clinical care while setting the stage for the implementation and evaluation of ChatGPT in clinical practice.

1.5 Innovation is required.

Experts suggest that the optimal application of AI in health care is not to replace human providers, but to improve the quality and efficiency of clinical practice by allowing clinicians to focus on areas where their combination of deep expertise, clinical experience, and direct personal contact lead to the greatest ‘value add’. As such, LLMs have the potential to radically enhance the accessibility of medical information for patients, while optimizing the clinical roles of healthcare professionals. In the setting of preoperative assessment, data from our team demonstrate that preoperative assessment by an anesthesiologist is associated with decreased mortality and hospital length of stay, as well as a greater number of days alive and at home after surgery, allowing decreased day of surgery cancellations, and ensuring efficient use of scarce OR resources.[2]

This study is innovative for 3 main reasons.

1. This will be the first study of its kind to assess the safety and efficacy of ChatGPT in answering patient-generated questions in the clinical setting, ensuring alignment with The Ottawa Hospital (TOH) strategic plan to ensure continuous improvements in perioperative medicine and to accelerate discovery at the national and international levels.
2. It will be the first study to compare the performance of ChatGPT in two different languages (French and English) which is a key parameter in a bilingual country such as Canada.
3. Finally, the outcome and the design of the study have been decided after engaging with patient partners of all sexes and genders through the hospital's Patient and Family Engagement Program ensuring that the findings of this study will be relevant and valuable not just for clinicians but more importantly to the patients that it affects.

This study is the first step in a program of research assessing the implementation of LLMs in the clinical setting. In this prospective observational study, we will explore if LLMs can provide non-inferior responses to those of anesthesiologists to patient queries. The next anticipated step will be to evaluate different LLMs and develop a feasibility study for each of them. Once we have established the clinical acceptability, utility, and feasibility of LLMs in preoperative anesthesiology care, we expect to engage in a clinical implementation study of LLMs.

2. OBJECTIVES

This study investigates whether AI Large Language Models (LLMs) like ChatGPT can effectively answer patients' preoperative anesthesia questions in a manner non-inferior to anesthesiologists, as judged by both patients and experts in preoperative assessment. If successful, we will be able to improve the efficiency and effectiveness of the preoperative assessments by optimizing the value added for each in-person interaction. If so, this has the potential to free up anesthesiologists to provide care in the operating room (OR) and help address surgical waitlists while maintaining an optimum level of care for patients.

Specifically, the hypotheses that we will test in our first-of-its-kind proposed prospective, single-center observational cross-sectional study will be to:

- Estimate ChatGPT non-inferiority to clinician's responses to patient queries in terms of gain in knowledge, actionable information, completeness, patient satisfaction, and perceived empathy in clinical settings.
- Estimate the incidence of hallucinations (plausible sounding but incorrect or nonsensical answers) by ChatGPT in responding to anesthesia-related queries.
- Compare the performance of ChatGPT across population sub-groups (English and French, different sex and gender, age sub-groups as defined by World Health Organization (WHO), and education level as defined by International Standard Classification of Education 11 (ISCED) of The United Nations Educational, Scientific and Cultural Organization (UNESCO).
- Engage and inform patients and key health system users via Integrated knowledge translation (iKT) throughout the research process.

3. PROPOSED TRIAL

3.1 Study design

To test the performance of ChatGPT in the clinical setting, a prospective, single center observational cross-sectional study using a non-inferiority design, will be conducted with patients from TOH in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines. The results of the study will be reported using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline.[18]

3.2 Patient and anesthesiologist selection

Following Research Ethics Board (REB) approval, we will recruit patients scheduled for a pre-surgery consultation with a physician anesthesiologist at the TOH PAU. Eligible patients who consent will provide baseline data as detailed below. This information will be collected and confirmed by a research assistant over the phone before the PAU visit. Certain data points may also be retrieved from the patient's medical records and verified during the phone call, and subsequently entered into REDCap by the RA. After obtaining consent, patients will be asked to respond to the following query: "**What is the most important question you had hoped to ask your anesthesiologist today about your anesthesia care, pain management or the time immediately around surgery?'**. This will form the unit of analysis for our study. In addition to patient participants, we will also recruit anesthesiologists. This will allow responses to each participant query to be rated by the participant and by a clinical expert.

3.2.1 Patient selection

Inclusion Criteria:

Age ≥ 18 years.

Elective non-cardiac surgery.

In-person preoperative anesthesiology consultation.

Willingness and ability to provide informed consent independently.

Exclusion Criteria:

Unable to communicate in English or French

Non-physician assessment.

3.2.2 Anesthesiologist selection

Inclusion Criteria:

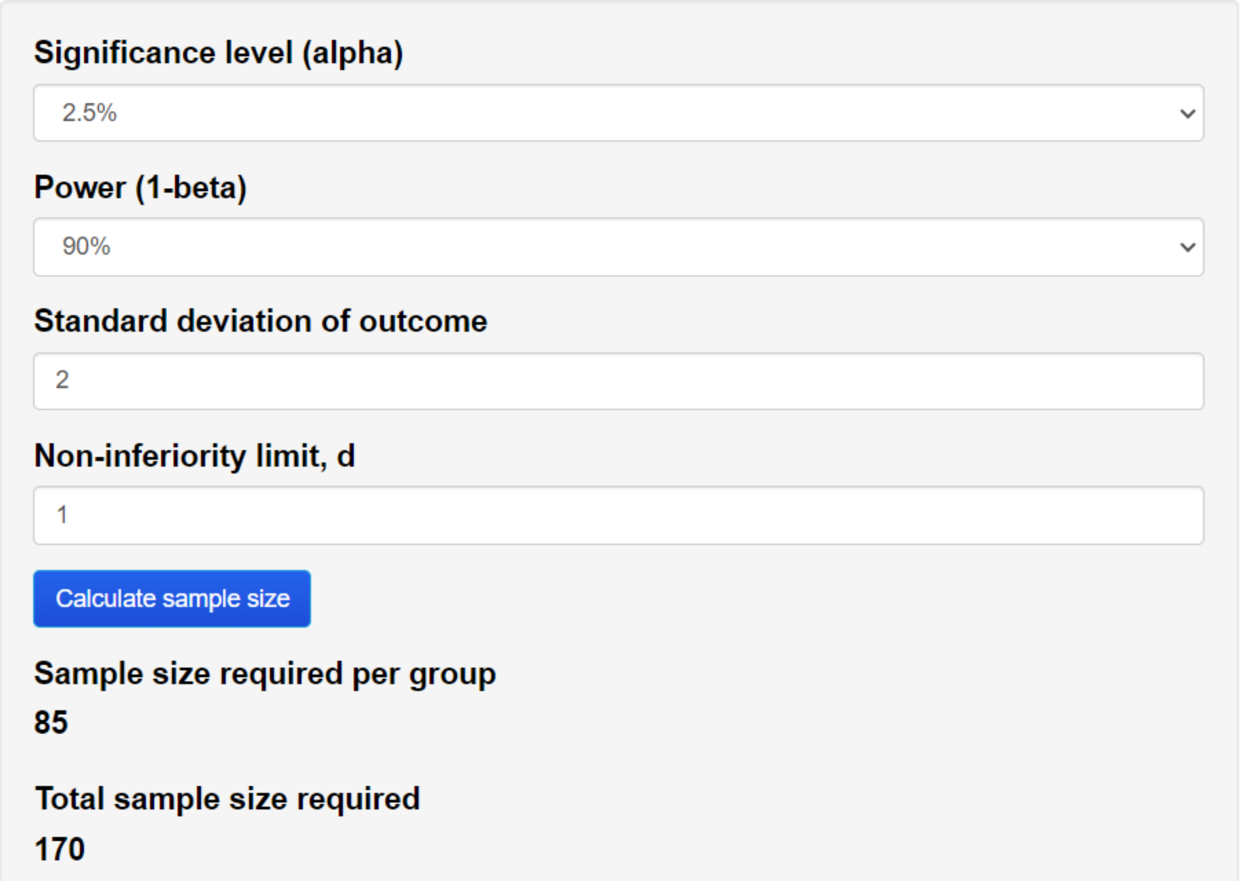
Anesthesiology residents or staff providing service at the PAU at TOH

Exclusion Criteria:

Anesthesiology residents or staff unwilling to participate in the research study.

3.3 Sample size

Minimal data are available to estimate expected knowledge changes with the implementation of LLMs in the perioperative setting. We based our effect size estimate on the related concept of patient feedback. With prior data from preoperative patients, we expect a SD of 2-points on the 11-point Likert scale and assume a 1 point minimally important difference to define our non-inferiority margin.[19]



Significance level (alpha)

2.5%

Power (1-beta)

90%

Standard deviation of outcome

2

Non-inferiority limit, d

1

Sample size required per group

85

Total sample size required

170

We would conclude that responses from ChatGPT are non-inferior to anesthesiologist responses if the lower bound of the 97.5% confidence interval is greater than or equal to -1, which will require a minimum n=170 to provide 90% power (1-sided alpha=0.025 for non-inferiority). To account for up to 10% attrition, we will recruit 190 total participants. We expect to recruit 8 patients/month/per center over 10 months. As we have two raters (i.e., patient and expert), to be judged completely non-inferior we will assess non-inferiority on the part of both raters.

4. STUDY PROCEDURES

4.1 Patients Identification, Consent, and query generation.

Patient participants will be identified using the Pre-Admission Unit schedule. The TOH PAU is now centralized with an office space at 1081 Carling Avenue, meaning that patients having surgery at the General, Civic or Riverside campus who require an in-person PAU assessment will visit this location.

Patients who are 18 years of age and older, who have an in-person PAU assessment scheduled, and who have permission for research contact documented in EPIC, will be phoned by a trained Research Assistant before their PAU visit.

- Over the phone, the Research Assistant will confirm with the patient that they have an in-person PAU assessment approaching. The Research Assistant will read the Verbal Consent Form to the patient and will obtain verbal consent from patients who are interested in participating.

- Once the patient provides verbal consent, the Research Assistant will ask the patient to provide and confirmed baseline data. Some data points may also be obtained from the patient's medical records and verified during the phone call, and subsequently entered into REDCap by the RA.

- Once consent and some baseline data are obtained, eligible patients will be asked to respond to the following query "**What is the most important question you hope to ask your anesthesiologist today about your anesthesia care, pain management or the time immediately around surgery?'** The question will be transcribed verbatim by the research assistant and read back over the phone to confirm that the transcribed query reflects the patient's intended query. If requested by the patient, the transcribed question will be emailed using a TOH account and avoiding sending patient's full name and clinical information. The use of email communication will remain limited, and the risks associated with this type of communication will be explained in the consent process.

4.2 Anesthesiologist Identification and Consent

Specialty-trained anesthesiologists in independent practice within the pre-admission unit will be informed about the study through in-person meetings, departmental emails, and follow-up from the research team. They will have the opportunity to provide written informed consent to participate as query raters or response reviewers. Communication will take place either in person or securely via Microsoft Teams, depending on each physician's preference and availability.

4.3 Generating responses to the patient query.

For each patient query, two responses will be generated.

The LLM response will be generated using GPT4, via the ChatGPT interface. An ex-novo ChatGPT account will be created for the purpose of the study and to maintain privacy and consistency throughout the study. Specifically, and to ensure consistency, A standard priming prompt will be provided "**I want you to respond as an anesthesiologist assessing a patient before their surgery. I will provide details about the patient's planned surgery and relevant medical history. Please respond with a comprehensive, empathetic, and concise answer, appropriate for a patient to read, in 160 words or less.**[20]." GPT4 will then be provided with the relevant, non-identifying participant details (limited to sex, decile of age, presence of routinely assessed comorbidities that are present in >5% of surgical patients, whether surgery is for cancer or non-cancer reasons, and surgical specialty), and the participant's query. An example follows:

'A male in his 60's with hypertension is coming for a general surgery procedure that is not for cancer. Their query is, 'How will my pain be managed after surgery to make sure that I can have a quick recovery?'

The following custom instruction prompt will be added to ChatGPT '**Below you will find details about a patient's planned surgery and relevant medical history. Please respond with a comprehensive and empathetic answer in 160 words or less to their question that would be appropriate for a patient to read. Be concise.'**

ChatGPT history will be turned off during the usage and the history will be deleted at the end of the study to maintain participants privacy.

The expert clinician response will be similarly generated. The anesthesiologist will be given the following prompt, '**. Below you will find details about a patient's planned surgery and relevant medical history. Please respond with a comprehensive and empathetic answer in 160 words or less to their question that would be appropriate for a patient to read. Be concise**.' The anesthesiologist will then be provided the same stem to generate their response as was provided to ChatGPT.

Both responses will be exported into standard word processing software to save each response under a unique identifier.

5. DATA COLLECTION & OUTCOMES MEASUREMENT

5.1 Data Collection

All patient participants will receive two parking vouchers: one for the PAU visit and one for the day of surgery. A research assistant will assist patient participants in rating responses from ChatGPT and anesthesiologists while they are waiting for their PAU appointment using patient-facing outcome measures present in Table 3. Patients will be able to provide responses directly into the Electronic Data Capturing (EDC) system developed in REDCAP using a tablet.

In this study, anesthesiologists will be designated as either "response generators" or "response raters." Response generators will craft answers to patient queries, while response raters—anesthesiologists with over five years of clinical experience—will evaluate the quality of these responses. For each participating patient, responses from both GPT-4 and the anesthesiologist response generator will be assessed by the raters using MS Forms within 7 days after the anesthesia consultation. To preserve objectivity, raters (patient and anesthesiologist response rater) will be blinded to the source of each response, which will be presented in random order, and they will not be involved in response generation. Additionally, a separate team of expert anesthesiologists will scrutinize all responses to detect any instances of hallucination. Patients will also provide their own ratings for the responses generated by GPT-4 and the anesthesiologist. Following each rating and anesthesiologist will debrief with the patient unblind the origin of each response addresses any potential hallucinations.

5.1.1 Patient's baseline data

- Following informed consent, baseline variables in **Table 1** will be collected by the Research Assistant including demographics and patient perceived technology comfort [21] [22].

Table 1: Baseline patient data

Baseline patient data	
Demographic Data	
Age	
Sex: Male; Female; I prefer not to say	
Gender: Man; Woman; Trans man; Trans woman, Gender-fluid, Nonbinary, Two-spirit, I don't identify with any option provided, I prefer not to say, Other, specify	
Race: American Indian or Alaska Native, Asian, Black, or African	

American, Native Hawaiian or Other Pacific Islander, White or Caucasian, Unknown or Not Reported	
Education level: No schooling Early childhood education Primary education Lower secondary education Upper secondary education Post-secondary non-tertiary education Short-cycle tertiary education Bachelor’s or equivalent level Master’s or equivalent level Doctoral or equivalent level Not elsewhere classified	
Primary language: French, English Bilingual (French-English) Other, specify	
Perioperative Data	
Planned surgery date	
Type of surgery. Thoracic (lung, chest cavity, esophagus, upper stomach Urology (bladder, kidney, prostate) Colorectal (small and large intestines, rectum, anus) Hepatobiliary (liver, bile ducts, pancreas) Orthopedic Gynecology	

Neurology Breast surgery Spine surgery Plastic surgery Vascular (veins and arteries) Otolaryngology (ears, nose, throat) Other specify:	
Is your surgery for cancer	
Is your surgery planned as minimally invasive (laparoscopic/keyhole) or open	
Previous surgery: yes, no if yes specify	
Previous anesthesia: yes, no if yes specify type	
Previous anesthesia complications: yes, no if yes specify	
What is your primary health condition?	
What are your principal comorbidities? High blood pressure Arrhythmia Chronic heart failure Coronary artery disease Heart valve problems Pulmonary hypertension Shortness of breath Steroid use in the past month Ascites Metastatic cancer Diabetes Hypothyroid	

Hyperthyroid Vascular disease Dialysis Smoking COPD Lung infection Kidney infection Bowel inflammation Sleep apnea Need help with daily activities like dressing, bathing, house work.	
Technology Comfort	
What type of internet-connected device do you normally use? Desktop Laptop, Tablet (iPad), Smartphone, Other; specify:	
Please rate your confidence using internet or related technology for health-related activities. 11-point Likert-scale (0 – Not at all confident to 10 – Very confident).	
Please rate your confidence using a computer or related technology for health-related activities. How strongly you disagree or agree with the following statement: Digital technologies being introduced to the perioperative setting will enhance patient care. 11-point Likert-scale (0 – strongly disagree to 10 – strongly agree).	

5.1.2 Anesthesiologist’s baseline data

Following informed consent, baseline variables of the anesthesiologist reviewer **Table 2** will be collected by the Research Assistant including demographics and level of training/years of experience. Anesthesiologists will be asked to complete 3 questions adapted from the Computer Literacy Questionnaire, [21] [22] to evaluate their perceived competence and comfort with internet-enabled devices.

Table 2: Baseline anesthesiologist data

Baseline anesthesiologist data	
Demographic Data	
Age	
Sex: Male; Female; I prefer not to say	
Gender: Man; Woman; Trans man; Trans woman, Gender-fluid, Nonbinary, Two-spirit, I don't identify with any option provided, I prefer not to say, Other, specify	
Race: American Indian or Alaska Native, Asian, Black, or African American, Native Hawaiian or Other Pacific Islander, White or Caucasian, Unknown or Not Reported	
Primary language: French, English Bilingual (French English) Other, specify	
Training Data	
Participant Year of completion of graduate degree	
Participant Year of completion of residency degree	
Participant Fellowship training: Perioperative fellowship, other fellowship, specify, no fellowship training, I prefer not to say	
Technology Comfort	
What type of internet-connected device do you normally use? Desktop Laptop, Tablet (iPad), Smartphone, Other; specify:	
Please rate your confidence using internet or related technology for health-related activities. 11-point Likert-scale (0 – Not at all confident to 10 – Very confident).	
How strongly you disagree or	

agree with the following statement: Digital technologies being introduced to the perioperative setting will enhance patient care. 11-point Likert-scale (0 – strongly disagree to 10 – strongly agree).	

5.2 Outcomes Measurement

5.2.1 Patient-facing outcome measures

Currently, LLMs performance evaluation in clinical contexts lacks standard outcome measures. Therefore, our choice of outcome metrics reflect key conceptual needs identified by patients and clinicians, as well as healthcare outcome and decision frameworks including the Institute for Healthcare Information's (IHI)[23] Quadruple Aim,[24 25] the Patient Education Materials Assessment Tool (PEMAT)[26 27] and the International Patient Decision Aid Standards (IPDAS). [28 29]

a) **Primary Outcome:** using the IPDAS and PEMAT tools we have identified the decision quality construct of 'knowledge' as primary outcome.

Patient evaluation of LLMs vs. anesthesiologist responses: Patients will be blinded to the source of the two responses provided to their query (GPT-4 response; Anesthesiologist response) and will be asked to complete the same questionnaire after reading both written responses. The questionnaire will include an 11-point Likert-scale question for the primary outcome knowledge. (Question: To what extent does the response provided address the knowledge or information that you hoped to gain in asking your question? 0-not at all addressed; 10-fully addressed).

b) **Secondary Outcome:** From the IHI Quadruple Aim we have identified the experience domain construct of 'satisfaction' as our secondary outcome. Patient satisfaction will be assessed using a likelihood to recommend measurement based on a 11-point Likert scale as recommended by the Institute for Healthcare Improvement's Quadruple Aim measurement guide.[18] (Question: Thinking about the response you received to your question, how likely are you to recommend the response's provider to a family member or friend going for a similar surgery? 0-not at all likely; 10-extremely likely).

c) **Tertiary outcomes:** As this field of work is in its infancy, we have engaged with key groups and through triangulation with patients and clinical experts, have prioritized exploring the following concepts:

- *Health literacy* as measured by merging two key domains of the Health Literacy Questionnaire (HLQ) (having sufficient information to manage my health; understand health information enough to know what to do).[30] This abbreviated version of the HLQ will include 1 patient-centered question. Each response to each patient query will be rated by the patient participant 11-point Likert scales. (Question: Based on the response I received, I have sufficient information and I can make a better decision about my surgical journey. 0-not at all sufficient; 10-fully sufficient).
- *Perceived empathy* – Assessing the way dialogue systems create perceptions of empathy unveils a range of technological, psychological, and ethical considerations that merit greater scrutiny during LLMs evaluation. There is currently no widely accepted evaluation method for determining the degree of empathy that any given system possesses.[31] The degree of perceived empathy conveyed by a LLMs will be assessed in this study by the patient. (Question: Empathy is the ability to understand and share the feelings of another. Based on the response received, to what extent did you find this response to be empathetic? 0-not at all empathetic; 10-extremely empathetic).
- *Perceived completeness of the response* – (Question: Completeness is how well the response adheres to the specific request you made by asking the question. Based on the response received, to what extent did you find this response to be complete? 0-totally incomplete as significant parts are missing; 10-totally complete as all required aspects of the question were addressed).

All patients' outcomes for each rater group (and related scale anchors) are described below:

Table 3: Patient-facing outcome measures

Patient-facing outcome measure
Primary Outcome: Patient perceived gain in knowledge
Question: To what extent did the response provided addresses the knowledge or the information that you hoped to gain in asking this question?
Answer: 0-not at all addressed; 10-fully addressed
Secondary Outcome: Patient satisfaction
Question: Thinking about the quality of response you received to your question, how likely are you to recommend this type of response to a friend, family member or patient going for a similar surgery as yourself? Please rate with 0-not at all recommend and 10 recommend without reservation
Answer: 0-not at all recommend; 10-recommend without reservation
Tertiary outcomes
a) Health literacy as actionable information for the patient
Question: Based on the response received, I have sufficient information and I can make a better decision about my surgical journey.
Answer: 0 – not at all sufficient; 10 – sufficient information

b) Patient perceived empathy
Question: Empathy is the ability to understand and share the feelings of another. Based on the response received, to what extent did you find this response to be empathetic?
Answer: 0 – not at all empathetic; 10 –extremely empathetic
c) Patient perceived completeness
Question: Completeness is how well the response adheres to the specific request you made by asking the question. Based on the response received, to what extent did you find this response to be complete?
Answer: 0 – totally incomplete (significant parts are missing); 10 –totally complete (addresses all aspects of the question and provides additional information or context beyond what was expected)**.**

Unblinding & Debriefing

After patient participants have completed their blinded rating of the 2 responses, the Research Assistant will inform them which response was generated by a trained anesthesiologist. They will receive a paper copy of this response. If they have questions or concerns, they will be able to discuss them with the anesthesiologist in the pre-admission unit. If necessary, the PI or representative will contact the patient for debriefing.

5.2.2 Anesthesiologist reviewer facing outcome measures.

We will provide a brief 10-minute online or in-person session explaining the study procedures, especially for raters, to ensure a clear understanding of the evaluation criteria and process. This session is optional but recommended to ensure consistency across all participants. Anesthesiologist reviewers will review and complete the baseline anesthesiologist data before completing the anesthesiologist reviewer facing outcome measure listed below after reviewing each response (i.e., the ratings for the first response will be completed prior to seeing the second response). The following outcomes will be investigated:

- Anesthesiologist reviewer perceived gain in knowledge

Question: Based on your expertise, does this response adequately address the patient’s query? 0-not at all addressed; 10-fully addressed.

- Anesthesiologist reviewer perceived accuracy

Question: Based on your expertise, and the patient’s query, to what extent did you find this response to be complete? 0-completely inaccurate; 10-completely accurate

- Anesthesiologist reviewer perceived completeness

Question: Based on your expertise, and the patient’s query, how accurate is this response? 0-totally incomplete as significant parts are missing; 10-totally complete as all required aspects of the question were addressed.

- Anesthesiologist reviewer perceived empathy

Question: Empathy is the ability to understand and share the feelings of another. Based on your expertise, and the patient’s query, to what extent did you find this response to be empathetic? 0-not at all empathetic; 10-extremely empathetic

- Anesthesiologist reviewer perceived hallucinations

Question: Hallucinations are a phenomenon where outputs that are nonsensical or altogether inaccurate are created. Based on your expertise, and the patient’s query, to what extent do you believe this response represents a hallucination? 0-not at all a hallucination, all information was sensible and accurate; 10-completely hallucinated, all aspects of the response were nonsensical and inaccurate.

Table 4: Anesthesiologist reviewer facing outcome measure.

Anesthesiologist reviewer facing outcome measure
Anesthesiologist reviewer perceived gain in knowledge:
Question: Based on your expertise, does this response adequately address the patient’s query?
Answer: 0-not at all addressed; 10-fully addressed
Anesthesiologist reviewer perceived accuracy
Question: Based on your expertise, and the patient’s query, is this response accurate?
Answer: 0- completely incorrect; 10- completely correct
Anesthesiologist reviewer perceived *completeness*
Question: Based on your expertise, and the patient’s query, to what extent did you find this response to the patient’s query to be complete?
Answer: 0 – totally incomplete (significant parts are missing); 10 –totally complete (addresses all aspects of the question and provides additional information or context beyond what was expected)**.**
Anesthesiologist reviewer perceived empathy
Question: Empathy is the ability to understand and share the feelings of another. Based on your expertise, and the patient’s query, to what extent did you find this response to be empathetic?
Answer: 0 – not at all empathetic; 10 –extremely empathetic
Anesthesiologist reviewer perceived hallucinations
Question: Hallucinations are a phenomenon where outputs that are nonsensical or altogether inaccurate are created by Large Language Models such as ChatGPT. Based on your expertise, and the patient’s query, to what extent do you believe this response represents a hallucination?
Answer: 0 – absolutely a hallucination; 10 – not at all a hallucination.

6. STATISTICAL ANALYSIS

We will calculate group-level descriptive statistics for patients, expert respondents, and expert raters using appropriate summary measures. A linear regression model will be used to estimate the mean difference on the Likert scale comparing responses from ChatGPT to anesthesiologists. Paired responses for each participant will be accounted for using an exchangeable covariance matrix. Exploratory, pre-specified subgroup analyses will compare ChatGPT, and anesthesiologists based on patient language, age, gender, and education level.

Participants' queries will be analyzed using inductive thematic analysis[32] [33] to identify themes. Each query will be coded within identified themes in duplicate, with a senior author ensuring face and content validity. If >30 responses are available within a response theme, we will re-estimate a theme-specific non-inferiority subgroup analysis.

As a data quality assessment, two independent reviewers will rate some responses, and the measurement of agreement between the reviewers using the intraclass correlation coefficient statistic will be performed.

Nonsensical responses will be coded as likely hallucinations. To estimate the occurrence of hallucinations, we will report both the mean and SD of the hallucination rating scale, as well as dichotomizing the scale for scores >8/10 (very likely hallucinations). A 95% binomial confidence interval will be estimated, using the proportion of hallucination. If we identify no likely hallucinations, our zero hallucinations-based interval would have an upper bound of 1.5% using the on the 3/n formula for 0 event proportions.[34]

Interim Data Analysis

The interim analysis will evaluate the safety and preliminary efficacy of ChatGPT in responding to patients' queries compared to clinicians in a clinical setting, after enrolling approximately (50 participants) 26% of the planned sample size.

Data collected will include:

Demographic Characteristics: Age, gender, ethnicity, etc.

Safety Analysis: Incidence and severity of hallucinations (incorrect or nonsensical answers), Solutions provided and trends over time.

Efficacy Analysis: Primary and Secondary Endpoints

Data Integrity and Completeness: Assessment and reporting of any protocol deviations.

Based on the interim analysis, the Data Monitoring Committee will recommend modification of protocol, termination of trial, additional actions.

Safety Stopping Rules

Hallucination Incidence: If the incidence of hallucinations (incorrect or nonsensical responses) from ChatGPT exceeds 25% of the responses, the trial will be stopped immediately.
Unanticipated Safety Concerns: The trial may be halted for further evaluation if regulatory authorities identify unanticipated safety concerns related to ChatGPT.
Ethical Considerations: The trial may be terminated early if there are ethical concerns about participant safety, welfare, or the integrity of the study.

Communication and Reporting
Stakeholder Notification: In the event of an early trial stoppage, relevant stakeholders, including study participants, investigators, regulatory authorities, and ethics committees, will be promptly notified.
Reasons for Stopping: The reasons for early termination will be clearly documented and reported according to regulatory requirements and ethical standards.

7. DATA MANAGEMENT

7.1 Data Storage

The Master Subject Lists (patients and anesthesiologists) will contain identifiers and be housed on TOH SharePoint site. Data will be collected locally by the site personnel and entered via the MS Forms database. The information processed by ChatGPT are subject to the US laws and jurisdiction as ChatGPT is a property of OpenAI, Inc. a California based company. The data will be stored on TOH secure SharePoint network with database access limited to only those requiring access, such as the Investigators and research staff.

7.2 Record Retention

The study team will maintain adequate records to enable the conduct of the study to be fully documented and the study data to be subsequently verified. Research and source documents must follow the ICH E6(R2) ALCOAC principles: Attributable, Legible, Contemporaneous, Original, Accurate and Complete. The Investigator Site File (ISF) will contain the study’s essential documents. Study and source records must be stored according to local practices and retained for at least ten years following closure of the study with the REB.

7.3 Other Future Research

The data collected from this study may be used to inform and guide future research on the integration of AI technologies in clinical practice, particularly in improving patient communication and decision-making during pre-anesthesia consultations.

8. ETHICS

8.1 Research Ethics Board

Federal regulations, the International Council for Harmonization (ICH) guidelines and the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS2 (2022) require that approval be obtained from a Research Ethics Board (REB) before participation of human subjects in research studies. Before study onset, the protocol, informed consent, and any other written information regarding this study to be provided to the subjects, must be approved by an REB. Documentation of all REB approvals will be maintained by the site and the principal investigator.

8.2 Ethical Conduct of the Study

The study will be performed in accordance with the ethical principles that have their origin in the Declaration of Helsinki, TCPS2 (2022), ICH E6 (R2), and all applicable regulations.

8.3 Protocol Amendments

Any amendment to the protocol will be approved by the Q/PI and submitted to the REB for ethics approval prior to implementation.

9. SEX AND GENDER BASED CONSIDERATIONS

Both biological (sex) and socio-cultural (gender) constitute relevant sources of variation in a number of clinical and subclinical conditions, affecting risk factors, prevalence, age of onset, symptomatology manifestation, prognosis, biomarkers and treatment effectiveness.[35] Observed sex and gender differences in health and wellbeing are influenced by complex links between both biological and social-economic factors, which are often surrounded by confounding variables such as stigma, stereotypes, and the misrepresentation of data. Consequently, health research and practices can be entangled with sex and gender inequalities and biases.[36] In recent years, the social awareness of such biases has increased and they have become even more evident with the introduction of widespread advance in biomedical AI. Arguably AI technologies can be seen as acting as a double-edged sword; on one hand, algorithms can magnify and perpetuate existing sex and gender inequalities if they are developed without removing biases and confounding factors, alternatively, they have the potential to mitigate inequalities by effectively integrating sex and gender differences in healthcare if designed properly. However, the performance of LLMs in different genders are key parameters that have not been reported to date. There is a need of comprehensive assessment of the impact of sex and gender during the implementation of LLMs in clinical setting.

This protocol was drafted after engaging with patient partners from both sexes and genders. Relevant to all objectives, to ensure representativeness of our cohort, surgical procedures eligible for inclusion are relevant to both females and males. We will report all descriptive, outcome and trajectory data disaggregated by gender through use of subgroup and stratified analyses using SAGER guidelines.[37] The results of this study may allow the development of gender specific solutions during the implementation of LLMs in the clinical setting.

10. PATIENT ENGAGEMENT

This protocol was drafted after engaging with patient partners through the hospital's Patient and Family Engagement Program (PFEP). We chose the SPOR principle-based framework [38] to optimize collaborative partnerships between researchers and patients. The collaboration starts with a summary of the research projects (including compensation, acknowledgement, authorship, training, and support for patient partners and researchers), followed by focused learning session, and patient partner impact. They assist in the codeveloping of research objectives, refinement of the questionnaire and development of methods to engage stakeholders.

Patient engagement activities and team member expectations will be documented in a Terms of Reference Document. Our budget will include compensation for patient partners commonly 25 CAD per hour for a total of 12 hours a year and acknowledgement in paper and abstract will be granted. Our goal is to build a strong and sustainable relationship through transparency, commitment, respect, communication, feedback, and continuous evaluation over the duration of the research period.

11. SIGNIFICANCE/IMPORTANCE

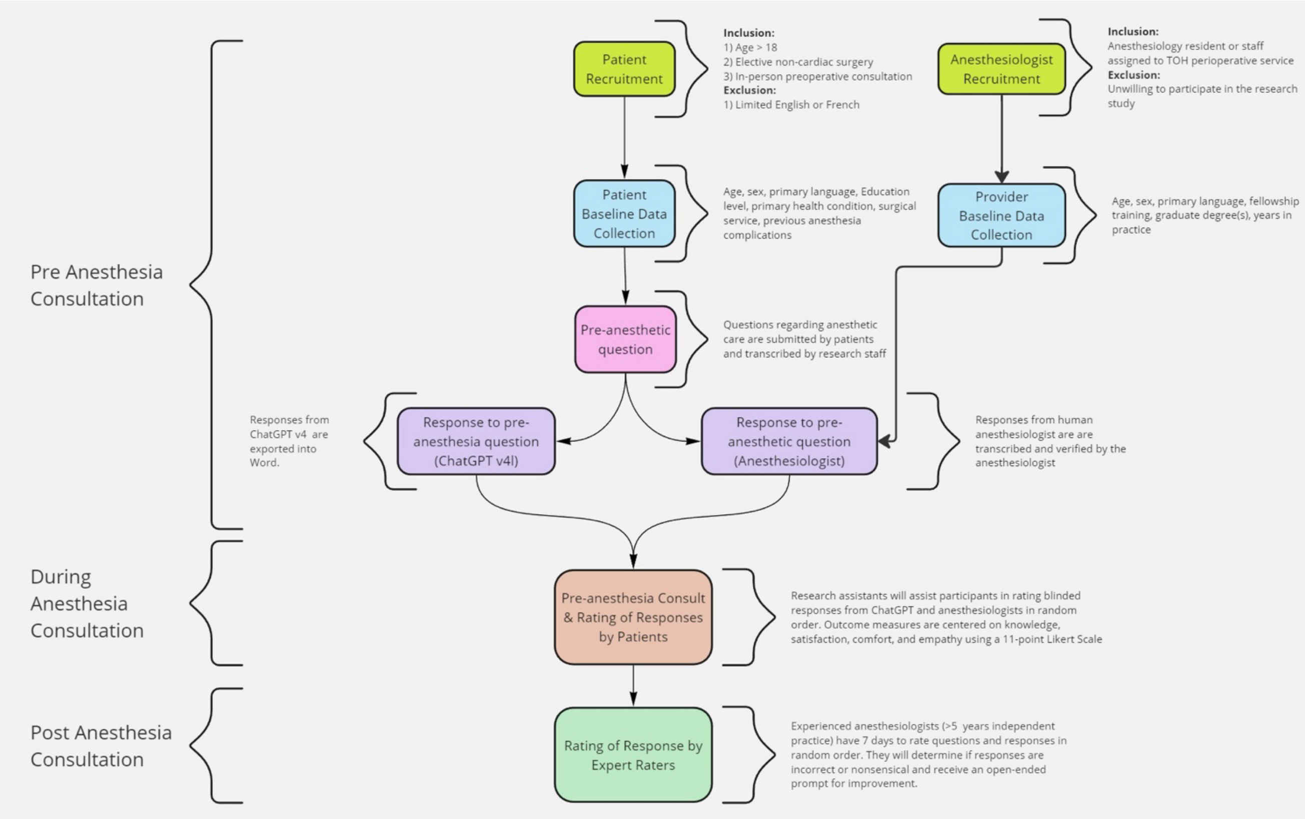
Given the increasing availability and performance of LLMs in combination with the fact that preoperative assessment is associated with decreased mortality and hospital length of stay, as well as a greater number of days alive and at home after surgery, the potential impact of perioperative implementation of LLMs will be significant. We have assembled an excellent team of clinicians and researchers with a combined expertise of in perioperative care, epidemiology, and AI. Collectively, our team is ideally positioned to address all of the objectives in this proposal and deliver robust evidence for supporting the effectiveness of implementation of LLMs in a patient-centric, and clinically relevant context for Canadian perioperative healthcare providers.

It is expected that the findings will contribute to the current literature on implementation of AI tools in the perioperative setting and inform patients, health care providers and policymakers with new insights into how to optimize health care delivery while maintaining high level of care for patients.

We worked with patients, system leaders and clinicians to identify the problem of anesthesiologist shortages. Once completed, we will disseminate our findings widely, using traditional methods like peer reviewed publications, and national and international conference presentations. We will also engage with implementation scientists and process engineers to understand how to best incorporate LLMs into routine perioperative patient flows.

If our research demonstrates that ChatGPT can provide responses to patients' queries prior to preoperative assessment in a manner that is non-inferior to answers from anesthesiologists, and without non-sense hallucinations, we will be able to increase the number of preoperative assessments performed by each anesthesiologist by optimizing the 'value add' of each in-person interaction. This will free up additional anesthesiologists to provide care in the operating room, allowing us to address surgical waitlists and optimize application of scarce and highly trained anesthesiologists to work at the top of their skill set in optimizing care for each surgical patient.

Figure 1. Study Flow


Pre Anesthesia Consultation
During Anesthesia Consultation
Post Anesthesia Consultation
Patient Recruitment
Inclusion:
1) Age > 18
2) Elective non-cardiac surgery
3) In-person preoperative consultation
Exclusion:
1) Limited English or French
Anesthesiologist Recruitment
Inclusion:
Anesthesiology resident or staff assigned to TOH perioperative service
Exclusion:
Unwilling to participate in the research study
Patient Baseline Data Collection
Age, sex, primary language, Education level, primary health condition, surgical service, previous anesthesia complications
Provider Baseline Data Collection
Age, sex, primary language, fellowship training, graduate degree(s), years in practice
Pre-anesthetic question
Questions regarding anesthetic care are submitted by patients and transcribed by research staff
Responses from ChatGPT v4 are exported into Word.
Response to pre-anesthesia question (ChatGPT v4l)
Response to pre-anesthetic question (Anesthesiologist)
Responses from human anesthesiologist are are transcribed and verified by the anesthesiologist
Pre-anesthesia Consult & Rating of Responses by Patients
Research assistants will assist participants in rating blinded responses from ChatGPT and anesthesiologists in random order. Outcome measures are centered on knowledge, satisfaction, comfort, and empathy using a 11-point Likert Scale
Rating of Response by Expert Raters
Experienced anesthesiologists (>5 years independent practice) have 7 days to rate questions and responses in random order. They will determine if responses are incorrect or nonsensical and receive an open-ended prompt for improvement.

REFERENCES

1. Henker R, Taki M. Challenges to Global Access to Anesthesia and Surgical Care. In: *Nurse Practitioners and Nurse Anesthetists: The Evolution of the Global Roles*. Springer; 2023:313-329.

2. Engel JS, Beckerleg W, Wijeysundera DN, et al. Association of preoperative anaesthesia consultation prior to elective noncardiac surgery with patient and health system outcomes: a population-based study. *British Journal of Anaesthesia*. Published online 2023.

3. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*. Published online 2018.

4. Gabashvili IS. The impact and applications of ChatGPT: a systematic review of literature reviews. *arXiv preprint arXiv:230518086*. Published online 2023.

5. Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:231211805*. Published online 2023.

6. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Published online 2018.

7. OpenAI. ChatGPT: Optimizing Language Models for Dialogue. Available online at: https://openai.com/blog/chatgpt/ (accessed June 29, 2023). Published online 2022. https://openai.com/blog/chatgpt/

8. Shahriar S, Hayawi K. Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. *arXiv preprint arXiv:230213817*. Published online 2023.

9. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. OpenAI. URL: https://openai.com/ product/gpt-4[. https://openai.com/ product/gpt-4[

10. Anil R, Dai AM, Firat O, et al. Palm 2 technical report. *arXiv preprint arXiv:230510403*. Published online 2023.

11. Nicholas DeNittis. What GPT-3 and AI-Generated Text Means for the Future of Written Content – with Peter Welinder of OpenAI. *https://emerj.com/ai-executive-guides/what-gpt-3-means-for-written-content/ Accessed: 13 Jan 2023.*

12. Birkett L, Fowler T, Pullen S. Performance of ChatGPT on a primary FRCA multiple choice question bank. *British Journal of Anaesthesia*. Published online 2023.

13. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health*. 2023;2(2):e0000198.

14. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Network Open*. 2023;6(10):e2336483-e2336483.

15. Yahagi M, Hiruta R, Miyauchi C, Tanaka S, Taguchi A, Yaguchi Y. Comparison of Conventional Anesthesia Nurse Education and an Artificial Intelligence Chatbot (ChatGPT) Intervention on Preoperative Anxiety: A Randomized Controlled Trial. *Journal of PeriAnesthesia Nursing*. Published online 2024.

16. Lim DY, Ke YH, Sng GG, Tung JY, Chai JX, Abdullah HR. Large language models in anaesthesiology: use of ChatGPT for American Society of Anesthesiologists physical status classification. *British Journal of Anaesthesia*. Published online 2023.

17. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*. Published online 2023.

18. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The Lancet*. 2007;370(9596):1453-1457.

19. Emily H, Yachnin D, Boland L, et al. Evaluation of a preoperative personalized risk communication tool: a prospective before-and-after study. *Canadian Journal of Anesthesia*. 2020;67(12):1749-1760.

20. Short CE, Short JC. The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights*. 2023;19:e00388.

21. Finch E, Hill AJ. Computer use by people with aphasia: A survey investigation. *Brain Impairment*. 2014;15(2):107-119.

22. Cartmill B, Wall LR, Ward EC, Hill AJ, Porceddu SV. Computer literacy and health locus of control as determinants for readiness and acceptability of telepractice in a head and neck cancer population. *International journal of telerehabilitation*. 2016;8(2):49.

23. Anderson JG, Aydin CE. Overview: Theoretical perspectives and methodologies for the evaluation of healthcare information systems. *Evaluating the organizational impact of healthcare information systems*. Published online 2005:5-29.

24. Stiefel M, Nolan K. A guide to measuring the triple aim: population health, experience of care, and per capita cost. IHI Innovation Series white paper. Available from URL: http://www.ihi.org/resources/Pages/IHIWhitePapers/AGuide toMeasuringTripleAim.aspx

25. Sikka R, Morath JM, Leape L. The quadruple aim: care, health, cost and meaning in work. *BMJ quality & safety*. 2015;24(10):608-610.

26. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient education and counseling*. 2014;96(3):395-403.

27. (2020) PEMAT for Printable Materials (PEMAT-P) | Agency for Healthcare Research and Quality. In: Agency Healthc. Res. Qual. https://www.ahrq.gov/health-literacy/patient-education/pemat-p.html. https://www.ahrq.gov/health-literacy/patient-education/pemat-p.html.

28. Sepucha KR, Borkhoff CM, Lally J, et al. Establishing the effectiveness of patient decision aids: key constructs and measurement instruments. *BMC medical informatics and decision making*. 2013;13(2):1-11.

29. Stacey D, Volk RJ, IPDAS Evidence Update Leads (Hilary Bekker KDS Tammy C Hoffmann, Kirsten McCaffery, Rachel Thompson, Richard Thomson, Lyndal Trevena, Trudy van der Weijden, and Holly Witteman). The international patient decision aid standards (ipdas) collaboration: evidence update 2.0. *Medical Decision Making*. 2021;41(7):729-733.

30. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). *BMC public health*. 2013;13(1):1-17.

31. Concannon S, Tomalin M. Measuring perceived empathy in dialogue systems. *AI & SOCIETY*. Published online 2023:1-15.

32. Braun V, Clarke V. *Thematic Analysis.* American Psychological Association; 2012.

33. Clarke V, Braun V. Thematic analysis: a practical guide. *Thematic Analysis*. Published online 2021:1-100.

34. Tuyl F, Gerlach R, Mengersen K. The rule of three, its variants and extensions. *International statistical review*. 2009;77(2):266-275.

35. Regitz-Zagrosek V. Sex and gender differences in health: Science & Society Series on Sex and Science. *EMBO reports*. 2012;13(7):596-603.

36. Hay K, McDougal L, Percival V, et al. Disrupting gender norms in health systems: making the case for change. *The Lancet*. 2019;393(10190):2535-2549.

37. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. *Research integrity and peer review*. 2016;1(1):1-9.

38. Greenhalgh T, Hinton L, Finlay T, et al. Frameworks for supporting patient and public involvement in research: systematic review and co-design pilot. *Health expectations*. 2019;22(4):785-801.