

Better Together: Exploring Patient Perspectives, Engagement, and Output Quality in Doctor-supervised Use of AI During Informed Consent Consultation with ChatGPT and Retrieval Augmented Generation (RAG)

Sascha Donner, Poroshista Knauer, Arne Kienzle, Jesse Dinneen, Joost Burger, Carsten Perka, Stefanie Donner

Submitted to: Journal of Medical Internet Research
on: April 07, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Better Together: Exploring Patient Perspectives, Engagement, and Output Quality in Doctor-supervised Use of AI During Informed Consent Consultation with ChatGPT and Retrieval Augmented Generation (RAG)

Sascha Donner^{1*} LLM; Poroshista Knauer^{2*} MD; Arne Kienzle² MD; Jesse Dinneen¹ PhD; Joost Burger² MD; Carsten Perka² MD; Stefanie Donner² MD, MHBA

¹Berlin School of Library and Information Science Humboldt-Universität zu Berlin Berlin DE

²Center for Musculoskeletal Surgery Charité - Universitätsmedizin Berlin Berlin DE

*these authors contributed equally

Corresponding Author:

Stefanie Donner MD, MHBA
Center for Musculoskeletal Surgery
Charité - Universitätsmedizin Berlin
Charitéplatz 1
Berlin
DE

Abstract

Background: Informed consent is essential for preoperative patient education, ensuring understanding of surgical procedures, risks, and postoperative care. AI chatbots, such as ChatGPT, offer potential to enhance patient education yet caution is needed.

Objective: To evaluate patient satisfaction and the quality of ChatGPT's responses in informed consent consultations for total hip arthroplasty (THA) and assess the impact of integrating subject- and hospital-specific data via Retrieval-Augmented Generation (RAG).

Methods: Patients scheduled for elective THA were assigned to three groups (n=12/group): traditional physician-led consultations (control), physician-assisted consultations using native ChatGPT, and physician-assisted consultations using ChatGPT+RAG. Patient satisfaction, preferences, health literacy, anxiety, and general attitude towards artificial intelligence were assessed using Likert scales and questionnaires. ChatGPT responses were evaluated for relevance, accuracy, clarity, completeness, and evidence-based content. Hallucinations were assessed and categorized for severity. Statistical analyses compared group differences.

Results: Patients using ChatGPT+RAG reported higher satisfaction with information delivery, informedness, and the amount of information received compared to native ChatGPT. AI-assisted groups asked more questions than the control group, with ChatGPT+RAG prompting the most engagement. Despite ChatGPT's benefits, patients predominantly preferred physician-led consultations, with state anxiety and patients' general attitude towards AI weakly influencing this preference. ChatGPT+RAG significantly reduced hallucinations and achieved superior scores across all quality metrics compared to native ChatGPT.

Conclusions: ChatGPT, especially with RAG integration, can enhance preoperative patient education by improving response quality and fostering engagement. However, patients continue to value human interaction, highlighting the complementary role of AI chatbots in informed consent.

(JMIR Preprints 07/04/2025:73717)

DOI: <https://doi.org/10.2196/preprints.73717>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to the public.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>, I will be able to make my accepted manuscript PDF available to anyone.

No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

1 Title

2 Better Together: Exploring Patient Perspectives, Engagement, and Output Quality in
3 Doctor-supervised Use of AI During Informed Consent Consultation with ChatGPT and
4 Retrieval Augmented Generation (RAG).

6 Authors

7 Sascha Donner^{*1} (sascha.donner@hu-berlin.de)

8 Poroshista Knauer^{*2} (poroshista.knauer@charite.de)

9 Arne Kienzle² (arne.kienzle@charite.de)

10 Jesse Dinneen¹ (jesse.dinneen@hu-berlin.de)

11 Joost Burger² (joost.burger@charite.de)

12 Carsten Perka² (carsten.perka@charite.de)

13 Stefanie Donner² (stefanie.donner@charite.de)

15 ^{*}Co-first authors

16 ¹ Berlin School of Library and Information Science, Humboldt-Universität zu Berlin,
17 Berlin, Germany

18 ² Center for Muskuloskeletal Surgery, Charité Universitätsmedizin Berlin, Berlin,
19 Germany

20 **Corresponding Author**

21 Stefanie Donner

22 Center for Muskuloskeletal Surgery, Charité Universitätsmedizin Berlin

23 Charitéplatz 1, 10117 Berlin

24 E-Mail: stefanie.donner@charite.de

25

26 **Disclosures**

27 None

28 **Abstract**

29 Introduction:

30 Informed consent is essential for preoperative patient education, ensuring
31 understanding of surgical procedures, risks, and postoperative care. AI chatbots, such
32 as ChatGPT, offer potential to enhance patient education yet caution is needed.

33

34 Objective:

35 To evaluate patient satisfaction and the quality of ChatGPT's responses in informed
36 consent consultations for total hip arthroplasty (THA) and assess the impact of
37 integrating subject- and hospital-specific data via Retrieval-Augmented Generation
38 (RAG).

39

40 Methods:

41 Patients scheduled for elective THA were assigned to three groups (n=12/group):
42 traditional physician-led consultations (control), physician-assisted consultations
43 using native ChatGPT, and physician-assisted consultations using ChatGPT+RAG.
44 Patient satisfaction, preferences, health literacy, anxiety, and general attitude towards
45 artificial intelligence were assessed using Likert scales and questionnaires. ChatGPT
46 responses were evaluated for relevance, accuracy, clarity, completeness, and
47 evidence-based content. Hallucinations were assessed and categorized for severity.
48 Statistical analyses compared group differences.

49

50 Results:

51 Patients using ChatGPT+RAG reported higher satisfaction with information delivery,
52 informedness, and the amount of information received compared to native ChatGPT.
53 AI-assisted groups asked more questions than the control group, with ChatGPT+RAG
54 prompting the most engagement. Despite ChatGPT's benefits, patients predominantly
55 preferred physician-led consultations, with state anxiety and patients' general
56 attitude towards AI weakly influencing this preference. ChatGPT+RAG significantly
57 reduced hallucinations and achieved superior scores across all quality metrics
58 compared to native ChatGPT.

59

60 Conclusion:

61 ChatGPT, especially with RAG integration, can enhance preoperative patient
62 education by improving response quality and fostering engagement. However,
63 patients continue to value human interaction, highlighting the complementary role of
64 AI chatbots in informed consent.

65

66 Keywords:

67 Large language model; GPT-4; AI chatbot; Hallucinations; Perioperative management;
68 Total hip arthroplasty; Patient education

69 1. Introduction

70 It is legally and ethically mandatory to provide patients with adequate information
71 during a preoperative surgical consent consultation. Preoperative education serves
72 not only as a vital communication tool between physician and patient but also holds
73 forensic importance. Time constraints in surgical practice often limit the amount of
74 time surgeons can dedicate to patients to inform them about the surgical procedure,
75 risks, and postoperative care during the information consent procedure. This can
76 leave patients feeling under-informed and anxious about their upcoming surgery.
77 With increasing constraints on medical staff resources and growing individual
78 workloads, there is also an increasing interest in boosting efficiency, particularly using
79 artificial intelligence (AI) [1,2].

80 Recent technological advancements have led to the development of large language
81 model-based tools such as the AI chatbot ChatGPT, which demonstrate improved text
82 analysis and generation capabilities that have the potential to greatly enhance patient
83 education and outcomes [3-6]. Patients can ask specific questions regarding their
84 planned surgical procedure and perioperative management and get in return answers
85 relevant to their questions with a possibly high level of accuracy. The potential of AI,
86 especially AI chatbots powered by generative pre-trained transformers (GPT) [7,8], is
87 increasingly recognized for its ability to transform patient engagement [9]. However,
88 the use of those tools also carries potential risks. Notably, the tools can produce
89 inaccurate, misleading, and potentially harmful responses, often referred to as
90 *hallucinations* [10]. Such hallucinations are expected to decrease with more advanced
91 versions like GPT-4o [11]. Particularly promising is the use of methods such as
92 retrieval-augmented generation (RAG), where the AI chatbot is provided with a
93 database containing additional, subject-specific information, though the effects of
94 such an approach on responses to patients' questions has not yet been examined.

95 Recent findings confirm that GPT-4 and GPT-4o exhibits exceptional accuracy and
96 efficiency in language and reasoning capabilities [12-15]. However, they also
97 emphasize the importance of developing more comprehensive benchmarks and
98 robust evaluation frameworks, incorporating qualitative assessments, human
99 judgment, and detailed error analysis [16,17]. This study therefore also aims at
100 investigating prompt-based assessment in the informed consent procedure.

101 Given its high degree of standardization and the development of robust postoperative
102 care protocols in recent years, hip arthroplasty serves as an exemplary model for
103 exploring the integration of AI into clinical procedures and patient education
104 frameworks [18-20]. Total hip arthroplasty (THA) is a procedure where a damaged or
105 worn-out hip joint is replaced with an artificial joint, particularly for patients with hip
106 osteoarthritis. This operation, widely regarded as the "operation of the century," is
107 noted for its profound impact on pain reduction and the enhancement of patients'
108 quality of life [21]. Before the operation, patients usually have questions concerning
109 the procedure and the physical rehabilitation process [22,23].

110 The process of information delivery is highly important and represents a central pillar
111 of patient education and the reduction of fear and anxiety [24,25]. Preoperative
112 anxiety is a common experience among patients, often stemming from inadequate
113 information about the surgical process. Therefore, patient education is of critical
114 importance prior to surgery. Evidence supports that better-prepared patients — those
115 who receive thorough preoperative education and physical conditioning — exhibit
116 improved postoperative outcomes, including accelerated recovery and reduced
117 complications. Therefore, enhanced patient education is an important component of
118 surgical care [26].

119 This is the first study to investigate surgical informed consent consultation using AI
120 chatbots prior to THA. Here, we analyzed ChatGPT's responses to patients' posed

questions preceding THA to assess its potential in preoperative patient education. Beyond the immediate benefits of patient education, the study explores the quality of the responses and potential hallucinations outputted by ChatGPT and systematically scrutinizes the system's current ability to deliver adequate responses. This study also examines patients' attitudes toward the use of an AI chatbot in a preoperative setting, along with potential cofactors influencing these attitudes. Assessing patient satisfaction with AI-driven consultations offers valuable insights into the acceptance and effectiveness of AI in patient interactions. The findings can inform the design of improved systems and the development of guidelines and protocols for integrating AI into patient care, particularly in sensitive areas like surgical informant consent.

Research questions:

1. What are patients' preferences regarding the use of an AI chatbot in conjunction with a physician in surgical informed consent in comparison to a physician only?
2. How does patient satisfaction compare across informed consent procedures involving only a physician, a physician and ChatGPT, or a physician and RAG-integrated ChatGPT?
3. How does the integration of RAG with ChatGPT improve the quality of responses to patients' questions compared to the native ChatGPT model?
4. How do patient sociodemographic characteristics and general attitude towards AI influence patients' preference towards using an AI chatbot in surgical informed consent?

2. Material and Methods

2.1. ChatGPT

In this study, we used the OpenAI ChatGPT version 4o (May 13, 2024). To ensure consistency, all questions were presented to ChatGPT in a single, continuous chat

146 session (per patient). The questions were asked by patients during their preoperative
147 surgical consultation before THA in our clinical practice. For the purposes of the study,
148 all patient queries were transcribed verbatim by the doctor into the ChatGPT
149 interface.

150 In our study we have used ChatGPT in two ways: (1) without any customization
151 (native) and (2) with customization via Retrieval-Augmented Generation (RAG),
152 described further below. The second approach allowed ChatGPT to access and
153 incorporate subject specific knowledge and thus potentially enhance the relevancy
154 and accuracy of its responses. The performance and output accuracy of the two
155 systems was compared as described below.

156 No patient-specific data was entered. The study was conducted in Germany, with all
157 interactions and materials presented exclusively in German.

158 **2.2. Recruitment**

159 Patients scheduled for elective, primary THA were recruited from the orthopedic
160 outpatient clinic or on the day of the information session. The study was conducted
161 from May to July 2024. The consent consultations were conducted by two physicians
162 who were pre-trained based on the prepared FAQ documentation.

163

164 Inclusion Criteria:

165 Age: Participants aged above 18 years

166 Diagnosis: Diagnosed with primary hip arthritis

167 Treatment History: Currently indicated for primary hip arthroplasty

168 Language: Able to read, understand, and complete study-related questionnaires in
169 German.

170 Exclusion Criteria:

171 1. Cognitive Impairment: Any cognitive impairment or psychiatric condition that
172 would hinder the participant's ability to provide informed consent and comply with
173 study procedures.

174 Prior operation: History of primary or revision arthroplasty on the contralateral side

175 The participants were provided with detailed written information about the aim of
176 the study, the study procedure, the inclusion and exclusion criteria, and the informed
177 consent about their participation in the study. The participants had the opportunity
178 to ask questions about any aspect of the information provided, the questionnaires,
179 and the study procedure.

180 **2.3. Study arms**

181 Each patient was assigned to one of the following study arms:

182 1. Control Group (human only): A physician conducting an informed consent
183 consultation in a traditional face-to-face setting.

184 Intervention Group I (native ChatGPT): ChatGPT was used in the Q&A session in
185 the presence of a physician.

Intervention Group II (ChatGPT+RAG): ChatGPT+RAG was used in the Q&A session in the presence of a physician.

From the total population of 36 patients, the first 12 patients were assigned to the Control Group. The next 12 patients were included in the native ChatGPT group, and the final 12 patients were allocated to the ChatGPT+RAG group.

In the RAG group an extensive internal document containing all the information about the surgical procedure and the perioperative management at our hospital was compiled and integrated into ChatGPT via RAG approach. This integration was achieved through the GPT App KeyMate.AI which adds the information from the document into the model's context window, effectively functioning as short-term memory. In the intervention group I, we added the following prompt text: „Please provide responses only in German and in the style of a medical doctor.”¹ In the intervention group II, we added a prompt text prior to each question posed to the AI chatbot: “Please provide responses only in German and in the style of a medical doctor. Please draw only on our internal FAQ document when responding.”² This setup is supposed to ensure that the AI chatbot responds preferably from the specific clinical information provided.

2.4. Data collection methods

Patient sociodemographic questionnaire and class-index (SDQ)

We collected the following patients' characteristics: age, gender, and highest level of education.

Patients' preferences and satisfaction questionnaire (PPSQ)

¹ in German: „Bitte antworte nur in deutscher Sprache und im Stil einer Ärztin oder eines Arztes.”

² in German: „Bitte antworte nur in deutscher Sprache und im Stil einer Ärztin oder eines Arztes. Bitte nutze zur Beantwortung ausschließlich die Informationen, die in unserem internen FAQ-Dokument enthalten sind.”

209 Data were collected using a self-constructed questionnaire that included 5 questions
210 on a 7-point Likert scale regarding the preference of a physician in the informed
211 consent procedure, the level of satisfaction of informedness, the procedure of
212 information transfer, the quantity of information as well as the preference for having
213 a physician even with less information.

214 Patients could also express their need for more information (yes/no).

215 Health literacy test (HLT) for Total Hip Arthroplasty

216 To ensure that patients were capable of giving informed consent and could in
217 principle understand the outputs of the AI chatbots, patients' health literacy was
218 measured with a questionnaire consisting of nine multiple choice questions. This
219 questionnaire was adapted from a questionnaire for total knee arthroplasty [27] but
220 with knee information and questions replaced by those for hip arthroplasty.

221 The General Attitudes towards Artificial Intelligence Scale (GAAIS)

222 The General Attitudes towards Artificial Intelligence Scale introduced by Schepman
223 and Rodway was evaluated to assess patients' attitude toward AI. [28] This 20-
224 question scale is divided into a positive (12 questions) and a negative (8 questions)
225 subscale. Both subscales capture emotions in line with their value. The 5-item scale
226 range from 1 to 5. The mean of the positive questions constitutes the overall score for
227 the positive subscale, while the same approach is applied to calculate the overall
228 score for the negative subscale. The higher the score, the more positive the attitude
229 towards AI. An overall scale mean was not recommended.

230

231 State-Trait-Anxiety Inventory (STAI)

232 To assess the actual state of anxiety (influenced by the current situation) and trait
233 anxiety (the general anxiety of the patient), the Spielberger State-Trait Anxiety

Inventory [29] was administered, specifically Grimm's abbreviated, validated version [30].

Output evaluation of AI chatbot responses

To assess how the quality of the AI chatbot responses, we conducted a detailed evaluation of the questions asked by patients during their interactions with the AI chatbot and the model's responses.

First, ChatGPT's output was evaluated based on Magruder et al. [31] for relevance, accuracy, clarity, completeness, and evidence based. Each criterion ranges from 1 to 5, with 5 being the best answer. We did not rate consistency between questions since every question was posed only once to the AI chatbot. For the accuracy criterion we compared the responses with the official guideline on indication criteria for total hip arthritis of the "German Society for Orthopedics and Trauma" (DGOU) [32], the FAQs of the German Arthroplasty Association (AE) [33], and our internal standard operating procedures. To evaluate the length of ChatGPT's output, we introduced an additional criterion that hold the balance between succinctness and informativeness. This criterion focused on if the response was appropriately concise while fully addressing the question without unnecessary elaboration. Scores ranged from 1 (overly brief or excessively length, missing necessary information, or including redundant content) to 5 (ideal length, providing a complete and focused answer that matched the question).

Second, we evaluated the level to which the AI chatbot produced so called hallucinations, i.e., "*content that is either nonsensical or unfaithful to the provided source content*" [34,35]. ChatGPT's output was classified as fully correct, partly hallucinated, and fully hallucinated. In addition, we annotated the level of hallucination using three *degrees*, following Rawte [36] and adapted for evaluation of AI chatbot outputs in the medical field. The degrees are *mild*, *moderate*, and *alarming*, with mild indicating minor hallucination that is superficial in terms of its

impact, moderate indicating a level of hallucination that introduces facts that are either fictitious or tangential to the topic at hand and which could cause the patient to become confused or feel insecure, and alarming indicating added information pieces that bear a radical dissemblance from the topic fed via the prompt or which could lead to serious harm to the patient.

2.5. Procedure

The study procedure and timing of data collection was identical for the control group and the two intervention groups up to timepoint 3, as follows (summarized in Figure 1 below):

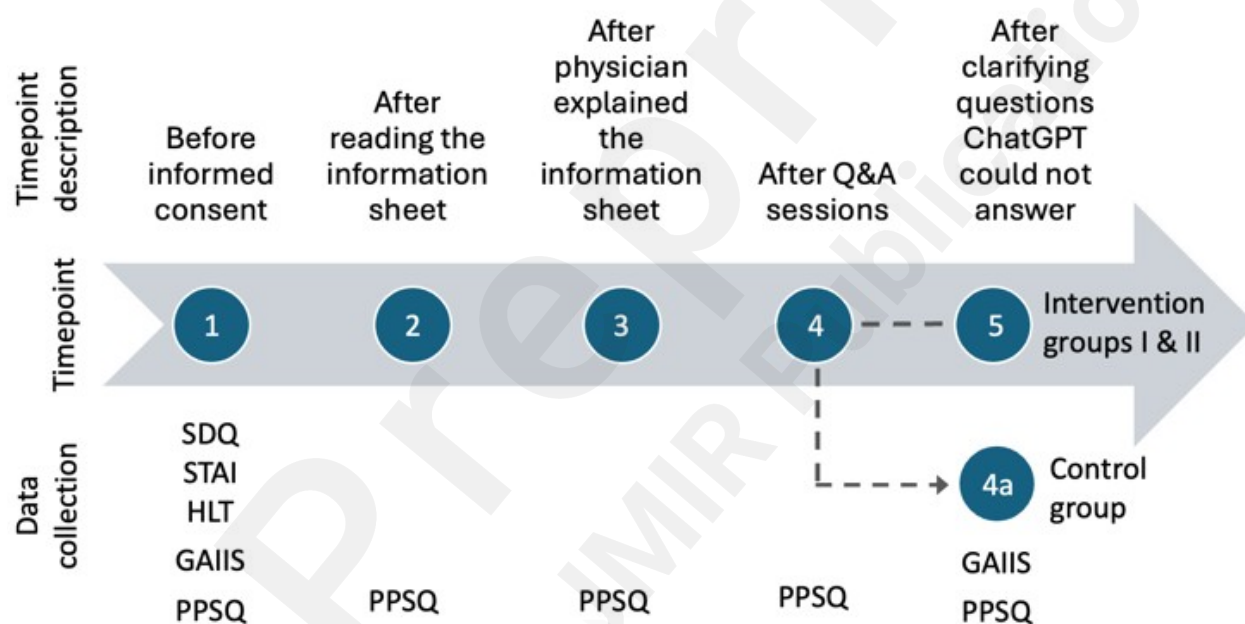
- Baseline Data Collection (timepoint 1): Initial data were collected prior to the informed consent procedure.
 - Information Sheet (timepoint 2): After the patient read the provided informed consent form.
 - Physician Explanation (timepoint 3): After the physician explained the operation, associated risks, and possible complications using the information consent form as a reference. At this point of time no questions were answered.
- From timepoint 4 onward, the procedures diverged:
- Control Group (timepoint 4): After patients asked questions in a conventional face-to-face setting and the physician answered without AI chatbot assistance.
 - Intervention Groups I and II (timepoint 4): After patients' questions were entered by the physician into the AI chatbot interface, and responses were read out word-for-word in German.

If chatbot responses were unclear or insufficient, the physician provided additional clarifications to ensure all questions were answered comprehensively and at a

satisfactory quality level (timepoint 5). In other words, the final data collection step for the Control Group was timepoint 4, whereas the Intervention Groups had a fifth data collection step, timepoint 5.

This structured process ensured that all patients, regardless of their group, received clear and sufficient answers to their questions during the informed consent procedure. All questions and answers of each information session were documented. In the following Figure 5, the course of the survey, the times of data collection, and method are shown.

Figure 1. Study procedure and time of data collection.



The PPSQ was completed 5 times (4 times in the human only group) during the informed consent procedure. The GAAIS was completed at the beginning and the end of the informed consent procedure, to evaluate if a change in the overall inclination towards AI has changed during the process. The SDQ, STAI, and HLT were completed prior to the informed consent procedure.

2.6. Statistical Methods

Sample Size Assumptions

301 In our exploratory study a sample size of 12 participants per group (one control group
302 and two intervention groups) was chosen based on the recommendations by Julious
303 et al. [37].

304 *Data Analysis*

305 Statistical analyses were performed using IBM SPSS Statistics (v. 29).

306 The Shapiro-Wilk test was applied to assess the normality of the data distribution for
307 continuous variables. The test revealed that none of the collected data were normally
308 distributed ($p < 0.05$). As a result, non-parametric statistical test was used for all
309 analyses.

310 A Friedman test was conducted to analyze differences of patients' preferences and
311 satisfaction levels (PPSQ) across the five time points: at the beginning of informed
312 consent, after reading the information sheet, after consulting the physician, after
313 using ChatGPT, and at the end of the informed consent process.

314 The Wilcoxon signed-rank test was applied to evaluate dependent pairwise
315 differences between satisfaction levels before and after the use of an AI chatbot. The
316 same test measured differences between The General Attitudes towards Artificial
317 Intelligence Scale (GAAIS) pre- and post-informed consent.

318 Spearman's-Rho was used to analyze correlation between GAAIS, State, Trait and
319 PPSQs.

320 The Mann-Whitney-U test was utilized to compare independent differences between
321 groups in the sociodemographic factors, the number of questions asked by
322 participants, the number of questions including hallucinations, and the quality of the
323 outputs of native ChatGPT and ChatGPT+RAG.

324 The Kruskal-Wallis test was employed to analyze the effects of highest educational
325 attainment on the outcomes of the patients' preferences and satisfaction levels
326 (PPSQ; i.e. across the seven independent variables).

327 Patient questions were categorized into five predefined groups: operation, postopera-
328 tive, complication, rehabilitation, and daily life. For the analysis, patient questions
329 from the two intervention groups, the native and the ChatGPT+RAG group, were dis-
330 tributed across these categories. The number of questions within each category was
331 counted for both groups. The Fisher-Freeman-Halton test was used to establish sig-
332 nificance in differences in the categorical data. Cramer's V was used for calculation of
333 effect size for Fisher-Freeman-Halton test; Friedmans effect size was calculated via
334 Kendall's ω , r was used for effect size of Mann-Whitney-U test and Wilcoxon signed-

335 rank test after the following formula: $r = \left| \frac{z}{\sqrt{N}} \right|$.

336 To assess inter-rater reliability (IRR) between the investigators in assessing the AI-
337 generated responses, Cohen's kappa coefficient was calculated. Kappa statistic values
338 were interpreted using the Landis and Koch criteria with values of 0.00 to 0.20
339 indicating slight agreement, 0.21 to 0.40 indicating fair agreement, 0.41 to 0.60
340 indicating moderate agreement, 0.61 to 0.80 indicating substantial agreement and
341 values of more than 0.80 indicating almost perfect agreement [38].

342

3. Results

3.1. Demographic data

Table 1. Demographic data.

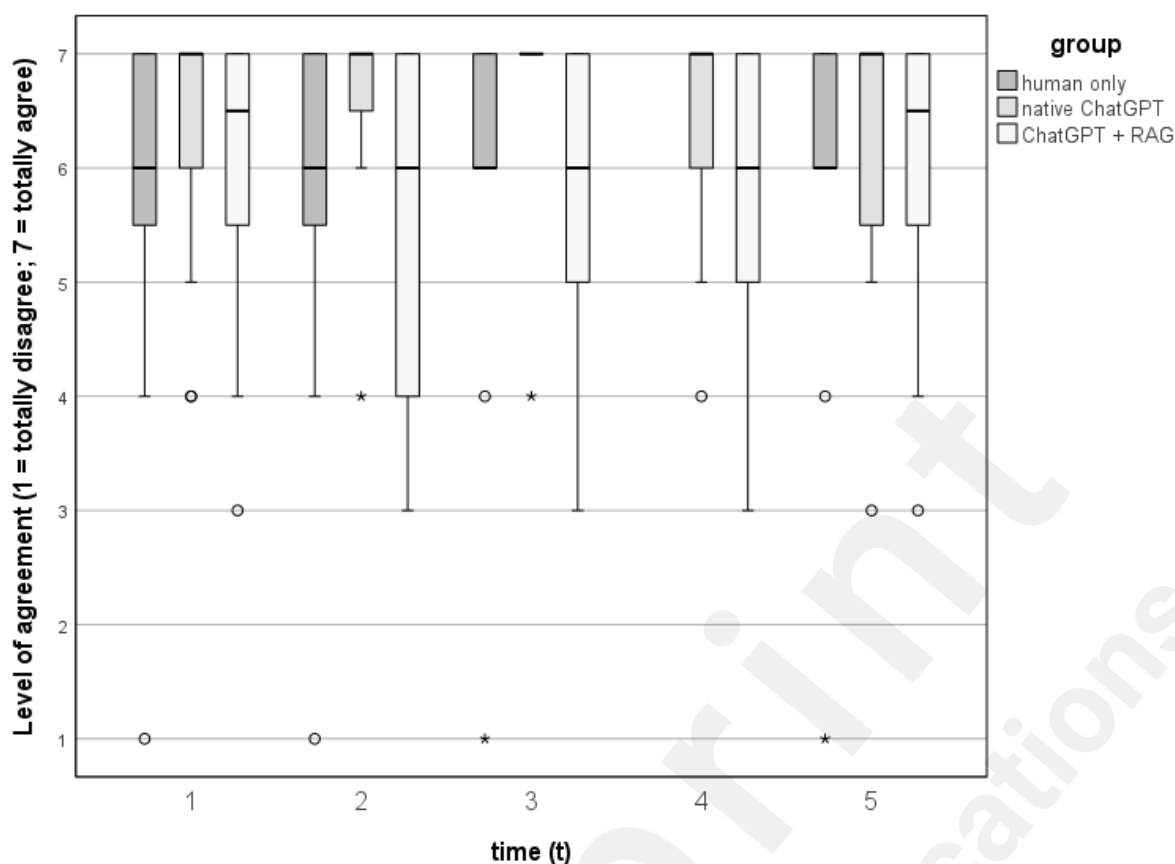
	Total	Control Group	Native ChatGPT	ChatGPT+RAG
Patients (n)				
	36	12	12	12
Age (years; mean (SD))				
	63.17 (8.84)	65.67 (8.40)	58.33 (9.49)	65.5 (8.75)
Age (years; min/max)				
	37/83	48/83	37/77	48/76
Gender (n; m/f)				
	23/13	5/7	10/2	8/4
Education: Vocational training (n/%)				
	20/55.56	6/50.00	8/66.67	6/50.00
Education: University degree (n/%)				
	16/44.44	6/50.00	4/33.33	6/50.00

3.2. Patients' preferences and satisfaction levels (PPSQ)

3.2.1 Need for a human doctor in the informed consent consultation

Average results are ranging between 5.50 and 6.65. Although slight shifts in patient's expressed need were observed during the study procedure (Figure 1), these changes were not statistically significant (human only: $p = 0.845$, $\omega = 0.227$; native ChatGPT: $p = 0.275$, $\omega = 0.107$; ChatGPT+RAG: $p = 0.048$, $\omega = 0.200$, with pairwise comparisons showing no differences across all groups).

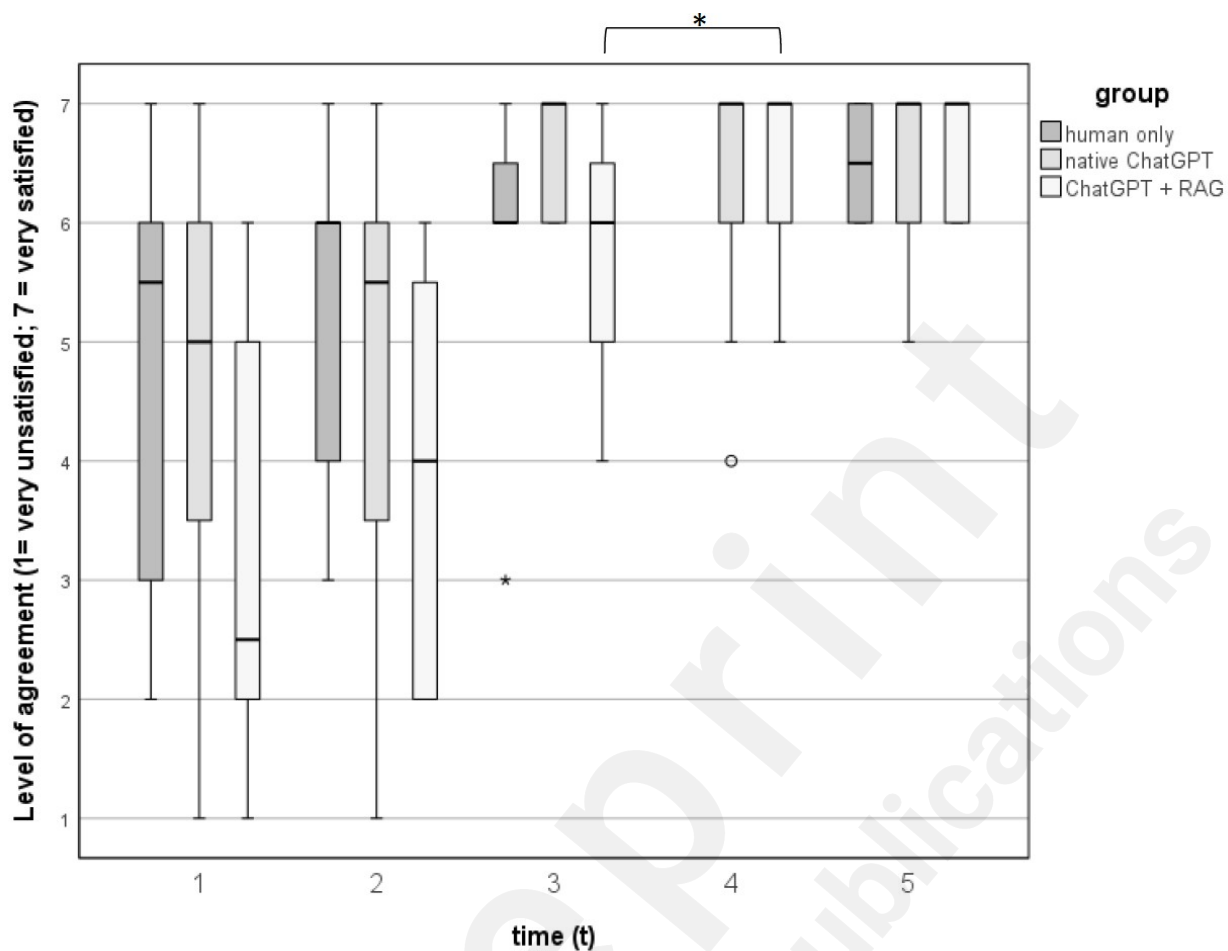
Figure 2. Comparison of patients' preference for human doctor during the informed consent consultation procedure.



3.2.2 Satisfaction with the process of information delivery

A difference in satisfaction levels was observed between the ChatGPT+RAG group and the native ChatGPT group (Figure 2). In the ChatGPT+RAG group, the patients were more satisfied with the process of information delivery after interacting with AI chatbot ($p = 0.014$, $r = 0.710$). Patients' satisfaction with the process of information delivery in the native ChatGPT-group do not differ after using AI chatbot compared to after the physician explained the THA operation procedure ($p = 0.276$, $r = 0.314$).

363 Figure 3. Satisfaction with the process of information delivery.



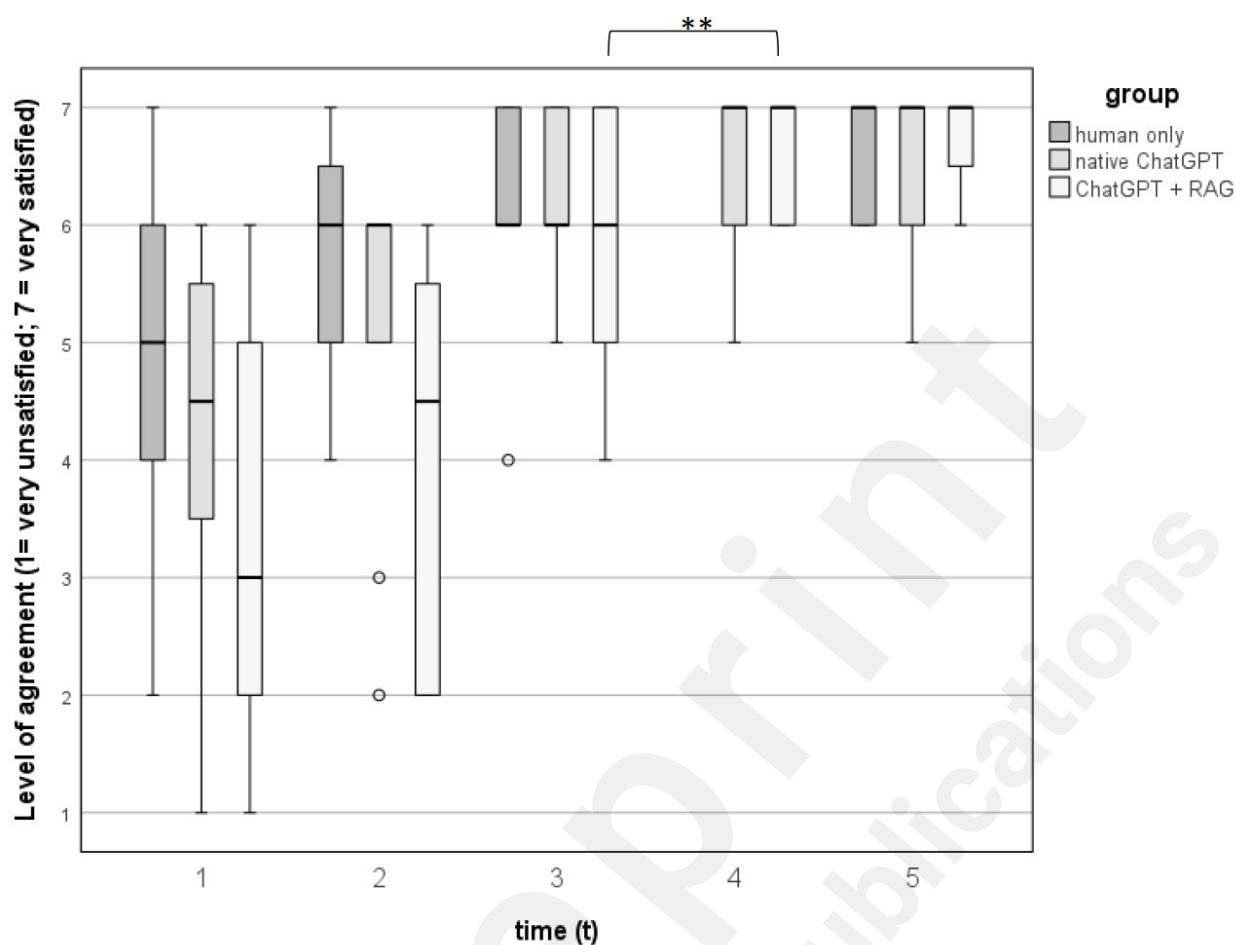
364

365 3.2.3 How satisfied are you with your perceived level of informedness?

366 A difference in satisfaction levels was observed between the ChatGPT+RAG group and
 367 the native ChatGPT group (Figure 3). In the ChatGPT+RAG group, the patients were
 368 more satisfied with their level of informedness after interacting with AI chatbot ($p =$
 369 0.007 , $r = 0.783$). Patients' satisfaction with level of informedness in the native
 370 ChatGPT-group do not differ after using AI chatbot compared to after the physician
 371 explained the THA operation procedure ($p = 0.059$, $r = 0.546$).

372

373 Figure 4. Satisfaction with perceived level of informedness.



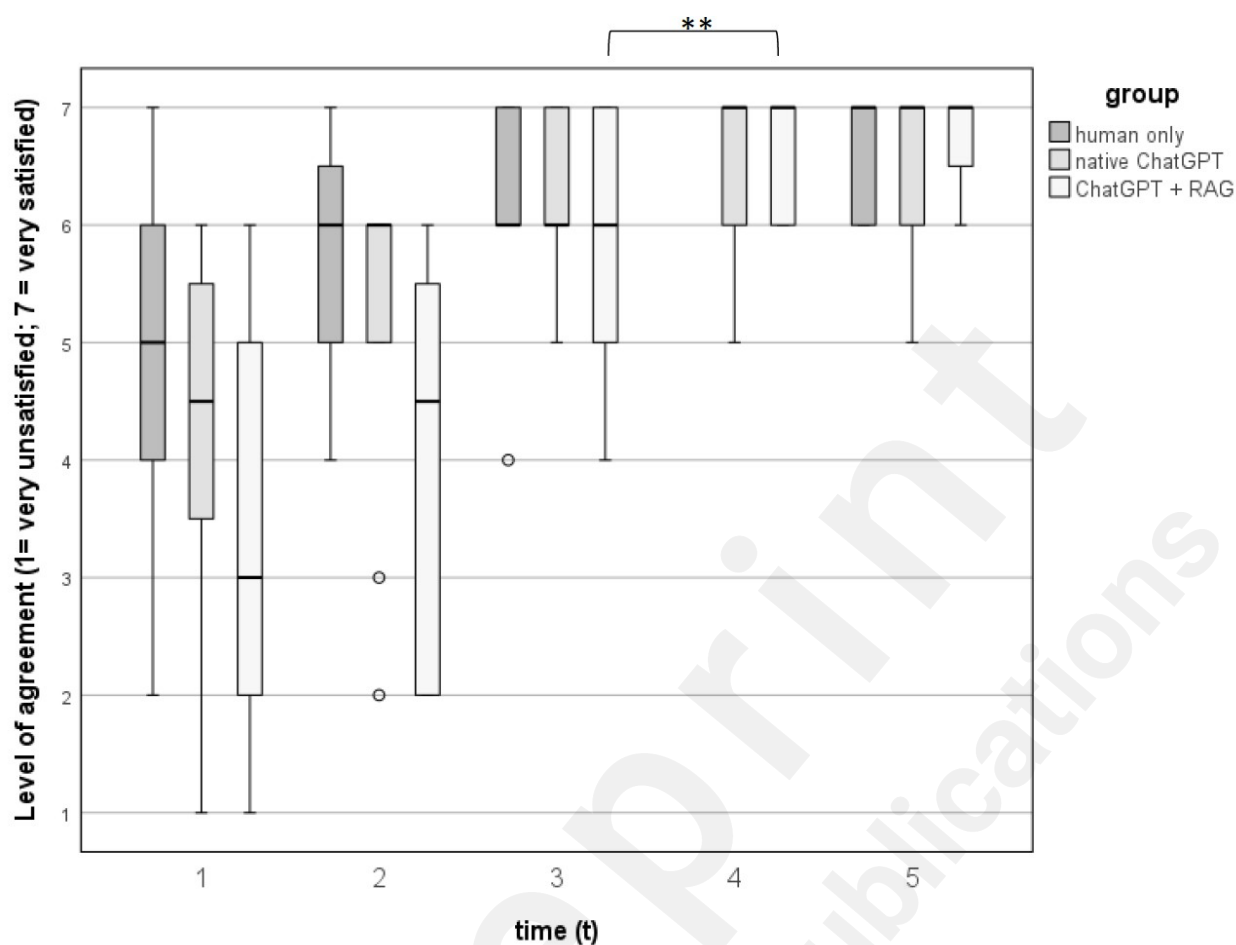
374

375 3.2.4 Satisfaction with the amount of information received

376 A difference in satisfaction levels with the quantity of information received was
 377 observed between the ChatGPT+RAG group and the native ChatGPT group (Figure 4).
 378 In the ChatGPT+RAG group, the patients were more satisfied with their level of
 379 received information after interacting with AI chatbot ($p = 0.021$, $r = 0.667$). Patients'
 380 satisfaction levels in the native ChatGPT-group do not differ after using AI chatbot
 381 compared to after the physician explained the THA operation procedure ($p = 0.317$, r
 382 $= 0.289$).

383

384 Figure 5. Satisfaction with amount of information received.



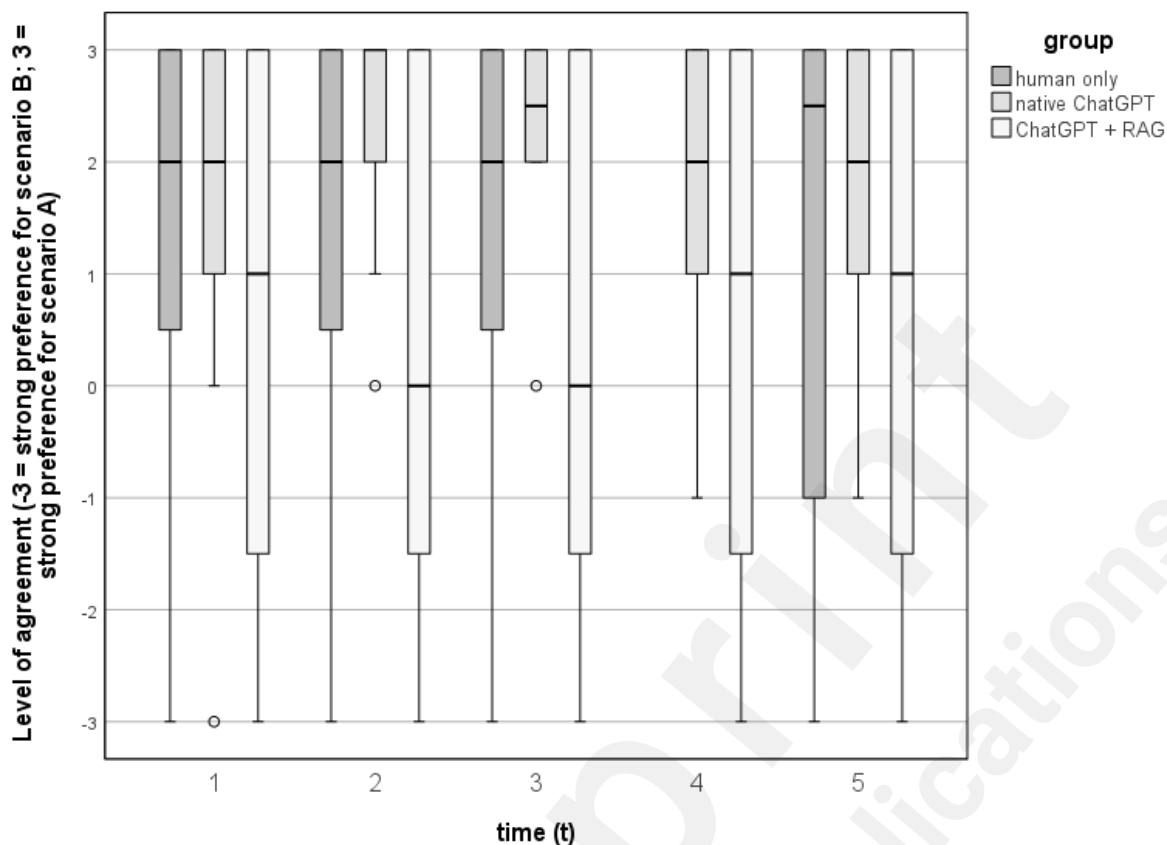
385

386 3.2.5 Preference for a human doctor vs AI chatbot

387 Results show a preference for scenario A, a human doctor with worse information-
 388 delivery vs. scenario B, AI-assistant only, with better information-delivery in all groups
 389 ranging from 1.25 to 1.42 in the human only group, from 1.75 to 2.33 in the native
 390 ChatGPT group, and from 0.58 to 0.83 in the ChatGPT+RAG group (Figure 5). The
 391 preferred scenario did not change during the procedure of informed consent, (human
 392 only $p = 0.753$, $\omega = 0.033$; native ChatGPT $p = 0.777$, $\omega = 0.037$; ChatGPT+RAG $p =$
 393 0.682 , $\omega = 0.031$).

394

395 Figure 6. Preference for a human doctor vs AI chatbot.



396 3.3. State-Trait-Anxiety Inventory (STAI)

397 The mean state anxiety among participants was 39.0% (SD: 17.6%), with values
398 ranging from 0% to 65.7%.

400 The mean trait anxiety among participants was 28.9% (SD: 14.4%), with values
401 ranging from 0% to 61.4%. There was no correlation between State, Trait, and the
402 preference of a physician in the informed consent procedure ($p_{\text{state}} = 0.158$; $p_{\text{trait}} =$
403 0.302), the level of satisfaction of informedness ($p_{\text{state}} = 0.522$; $p_{\text{trait}} = 0.605$), the
404 procedure of information transfer ($p_{\text{state}} = 0.856$; $p_{\text{trait}} = 0.822$), the quantity of
405 information ($p_{\text{state}} = 0.566$; $p_{\text{trait}} = 0.330$) as well as the preference for having a
406 physician even with less information ($p_{\text{state}} = 0.087$; $p_{\text{trait}} = 0.172$).

3.4. Sociodemographic factors' impact on PPSQ values

Sociodemographic factors overall did not significantly influence patients' attitudes toward the necessity of a physician in the informed consent process nor their preference for a human doctor versus an AI chatbot (i.e., scenario A vs. scenario B). However, age was a significant factor in satisfaction with the process of information delivery, with older patients (55% above the median age of 63.16 years) reporting higher satisfaction levels compared to younger patients (45%; $p = 0.046$, $r = 0.332$). Additionally, significant differences were observed in satisfaction levels regarding the amount of information provided, with male patients (64%) reporting higher satisfaction than female patients (36%; $p = 0.022$, $r = 0.383$). No significant effects were identified related to the patient's highest level of education.

3.5. Health literacy impact on PPSQ values

The health literacy test (HLT) had a mean of 7.92 right answered questions (SD 1.1), with a range from 6 to 9 right answers. No significant effects were identified related to the patients' health literacy test outcomes. No significant relationship between the HLT score and the PPSQ values was identified.

3.6. Attitude towards AI (GAAIS)

Patients' attitude towards AI did not change after using the AI chatbot in the informed consent consultation despite GAAIS negative in the native ChatGPT group ($p = 0.024$, $r = 0.650$; Table 2 and 3).

Table 2. Attitude towards AI towards AI before informed consent and after the procedure.

GAAIS	Pre-informed consent (t = 1)			Post-informed consent (t = 5)		
	Human only	Native ChatGPT	ChatGPT+RAG	Human only	Native ChatGPT	ChatGPT+RAG
Positive subscale						
	3.2 SD 0.7	3.6 SD 0.6	3.5 SD 0.6	3.1 SD 0.9	3.6 SD 0.6	3.4 SD 0.6
Negative subscale						
	2.7 SD 0.6	3.2 SD 0.7	3.2 SD 0.7	2.7 SD 0.6	3.4 SD 0.7	3.3 SD 0.7

Table 3. P-value for attitude towards AI pre vs. post informed consent GAAIS for positive and negative.

GAAIS	Human only	Native ChatGPT	ChatGPT+RAG
Pre vs. post positive			
	p = 0.182 r = 0.385	p = 1.0 r = 0	p = 0.833 r = 0.061
Pre vs. post negative			
	p = 0.865 r = 0.049	p = 0.024* r = 0.650	p = 0.066 r = 0.531

428 Significant correlation was shown between GAAIS positive and preference of a physi-
429 cian in the informed consent procedure ($p = 0.002$; $\rho = 0.502$) and between GAAIS
430 negative and preference of a physician in the informed consent procedure ($p = 0.002$;
431 $\rho = 0.500$).

432 Figure 7. Correlation between GAAIS positive and preference of a physician in the informed consent
433 procedure.

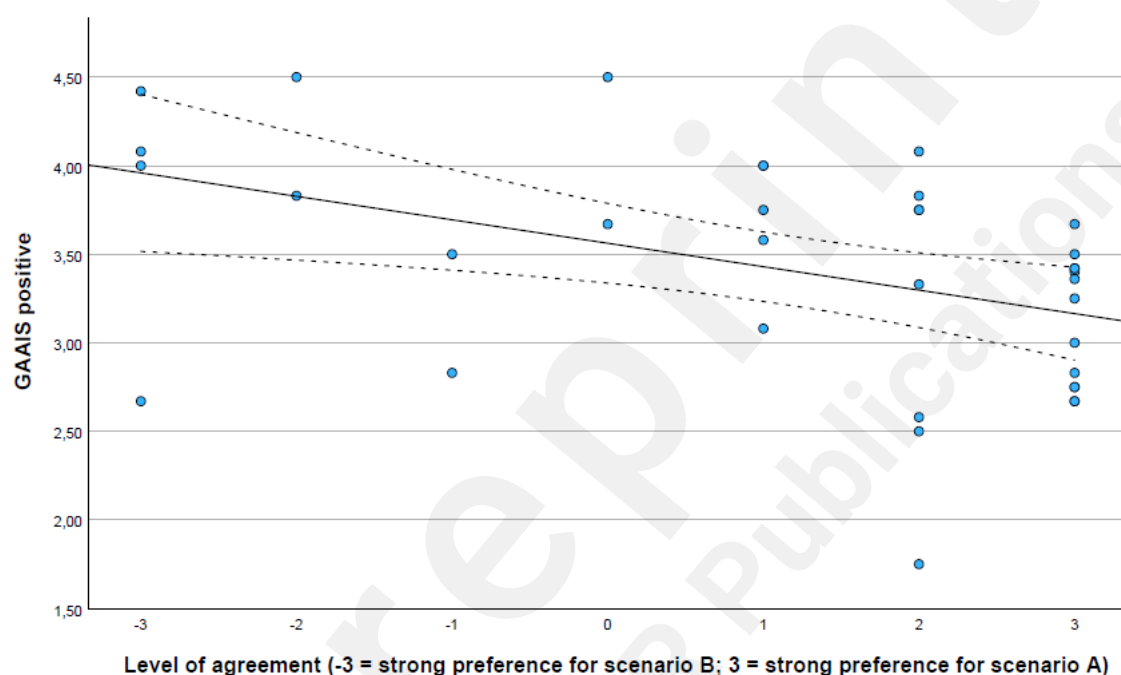
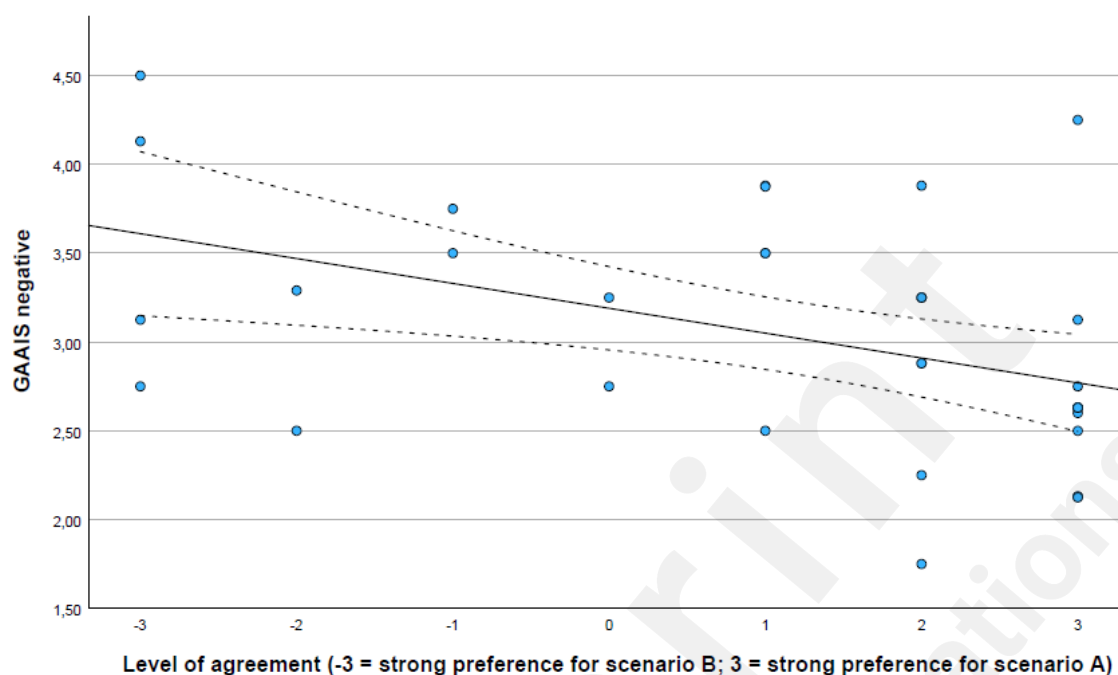


Figure 8. Correlation between GAAIS negative and preference of a physician in the informed consent procedure.



3.7. Patient questions and AI chatbot responses

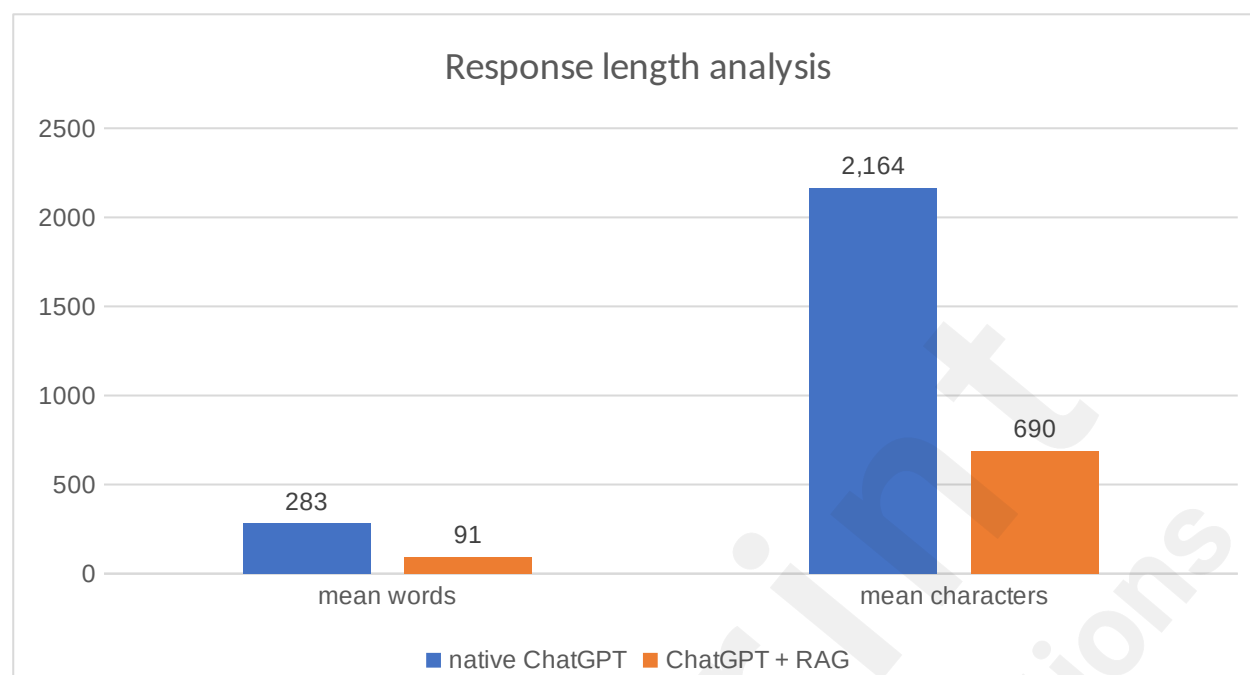
Examining the number of questions asked per group (Table 3), the control group stood out with fewer questions, totaling only 20, compared to the RAG group, which asked 52 questions ($p = 0.002$, $r = 0.621$). The native ChatGPT group asked 39 questions, which was not statistically significant compared to the human only group ($p = 0.056$, $r = 0.390$). Additionally, no difference was observed between the RAG and native ChatGPT groups ($p = 0.219$, $r = 0.251$).

Table 4. Number of questions asked by patients in each of the three groups.

	Human only	Native ChatGPT	ChatGPT+RAG
Total (n)			
	20	39	52
Mean			
	1.67	3.25	4.33
Min (n)			
	1	1	1
Max (n)			
	4	7	7

The average length of responses was measured for both intervention groups. In the native ChatGPT group, responses averaged 283.1 words and 2,163.6 characters while in the ChatGPT+RAG group, responses averaged 91.4 words and 689.8 characters (Figure 6). This represents a ratio of 3.098 for word count ($p < 0.001$, $r = 0.784$) and 3.136 for character count ($p < 0.001$, $r = 0.781$), indicating that native ChatGPT generated outputs are approximately three times longer than ChatGPT + RAG.

452 Figure 9. Length of ChatGPT responses.

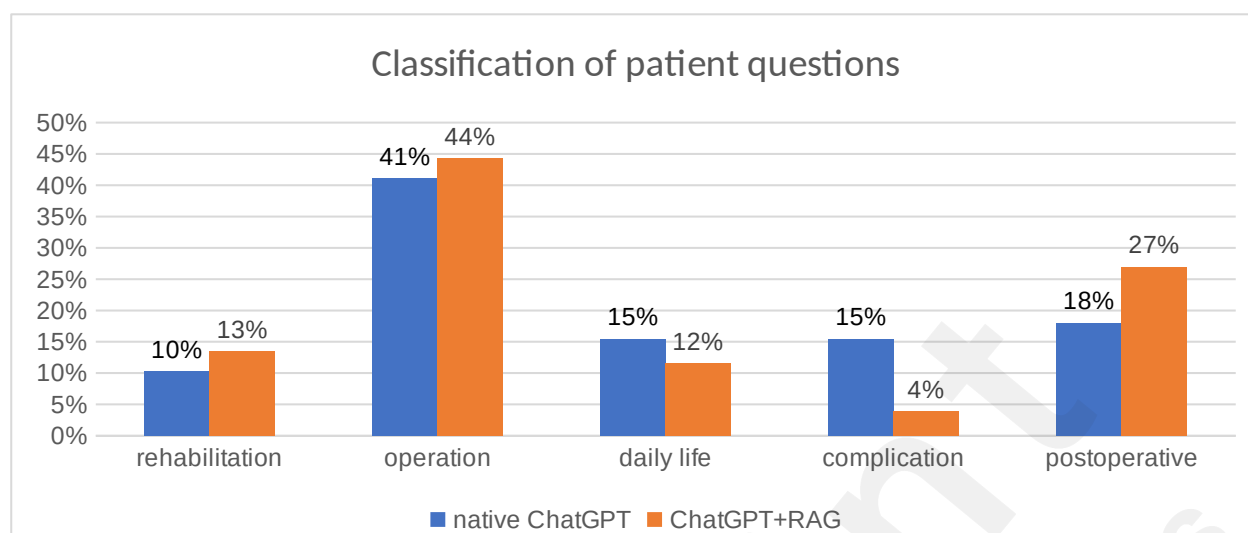


453

454 Figure 7 shows, that in both, the native ChatGPT and ChatGPT+RAG groups, most
455 patient questions were focused on the operation itself, followed by questions related
456 to postoperative management. Fewer questions were asked about rehabilitation,
457 daily life, and complications, highlighting a primary interest in the surgical procedure
458 and immediately afterwards. No differences were observed between the two groups
459 in the distribution of question classifications ($p = 0.350$, $V = 0.226$).

460

461 Figure 10. Classification of patient questions.



462

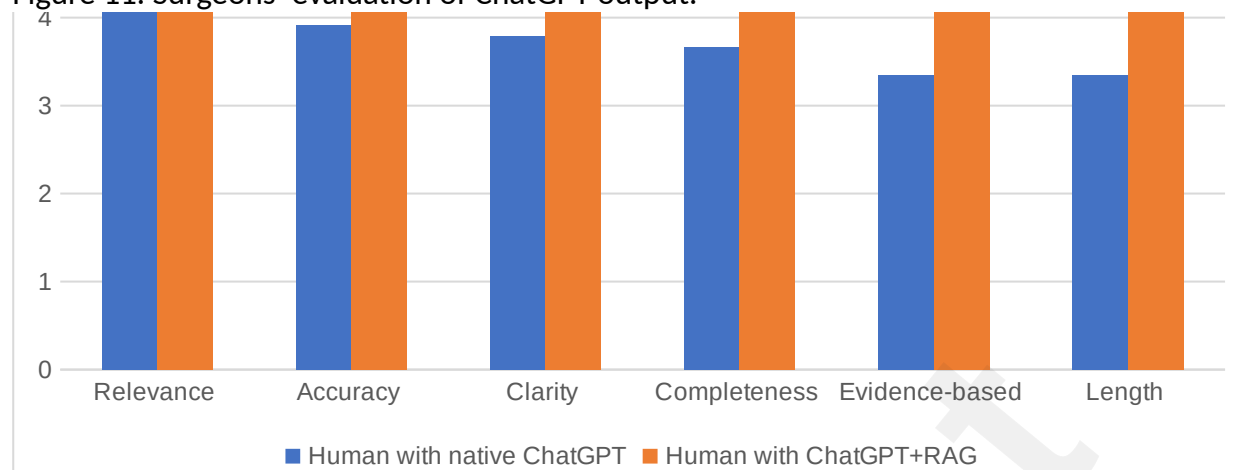
463

464 3.8. Evaluation of the ChatGPT output quality

465 All together 91 questions were prompted and the output was evaluated. Figure 8
 466 shows the results of physicians' evaluation of the ChatGPT outputs for every category
 467 (relevance, accuracy, clarity, completeness, evidence-based, and the length of the
 468 output). ChatGPT+RAG had better outcomes in every category compared to native
 469 ChatGPT ($p < 0.001$; $r(\text{relevance}) = 0.504$; $r(\text{accuracy}) = 0.365$; $r(\text{clarity}) = 0.565$;
 470 $r(\text{completeness}) = 0.479$; $r(\text{evidence-based}) = 0.483$; $r(\text{length}) = 0.726$). The minimum
 471 value for ChatGPT+RAG was 4.625, while its maximum value was 4.961. The minimum
 472 value for native ChatGPT was 3.346, while its maximum value was 4.115. Across all
 473 outputs, the kappa value was 0.761 (asymptotic standard error = 0.032, $t = 22.705$, p
 474 < 0.001), indicating a substantial level of agreement.

475

Figure 11. Surgeons' evaluation of ChatGPT output.



476

477 ** significant at the 0.01 level (2-tailed).

478 * significant at the 0.05 level (2-tailed).

479 We observed significantly less hallucinations in the RAG group vs. in the native
480 ChatGPT group (9.6% vs.38.5%). (See Table 5). For four out of the five hallucinated
481 responses in the ChatGPT+RAG group, no information was available in the FAQ
482 documentation provided to ChatGPT via RAG from which the AI chatbot could draw.

483

484 Table 5. Hallucinations in ChatGPT responses.

	Total	Native ChatGPT	ChatGPT+RAG
Questions asked (n, (%))	91 (100%)	39 (100%)	52 (100%)
Not hallucinated (n, (%))	71 (78.0%)	24 (61.5%)	47 (90.4%)
Hallucinated (n, (%))	20(22.0%)	15 (38.5%)	5 (9.6%)
Partly hallucinated (n, (%))	15 (16.5%)	12 (30.8%)	3 (5.8%)
Fully hallucinated (n, (%))	5 (5.5%)	3 (7.7%)	2 (3.8%)
Mild severity (n, (%))	8 (40.0%)	7 (46.7%)	1 (20.0%)
Moderate severity (n, (%))	9 (45.0%)	7 (46.7%)	2 (40.0%)
Alarming severity (n, (%))	3 (15.0%)	1 (6.6%)	2 (40.0%)

485

486 ChatGPT+RAG has a significant lower hallucination score compared to native ChatGPT
 487 ($p = 0.002$ $r = 0.332$). The degree of severity shows no significant difference ($p = 0.168$
 488 $r = 0.340$).

489 Examples for *mild*, *moderate*, and *alarming* hallucinations identified in our study were
 490 as follows:

491

492 Textbox 1. Example - mild hallucination.

493 Patient question: "I am having a THA, how can I sleep after the operation?"

494 Excerpt from the answer: "1. supine position: This is the preferred sleeping
495 position immediately after the operation. Place a pillow between your legs to
496 keep the hip in a stable position and prevent the operated leg from rotating
497 inwards or outwards..."

498 Explanation for categorization: Due to today's prevailing minimally invasive
499 surgical techniques, a cushion between the legs in supine position is no longer
500 common or necessary.

501
502 Textbox 2. Example - moderate hallucination.

503 Patient question: "Which type of THA does squeak?"

504 Excerpt from the answer: "...ceramic-on-polyethylene bearing: This
505 combination may also occasionally squeak, although this is less common than
506 with pure ceramic-on-ceramic dentures..."

507 Explanation for categorization: Under normal conditions, this bearing pair
508 should not squeak. If this pair of bearings squeaks, the reason for the
509 squeaking must be analyzed.

510
511 Textbox 3. Example – alarming hallucination.

512 Patient question: What do you do with a dislocated hip?

513 Excerpt from the answer: "After reduction, an X-ray or MRI is performed to
514 ensure that the hip is positioned correctly and to rule out any associated
515 injuries such as fractures or damage to ligaments and soft tissue..."

516 Explanation for categorization: An X-ray must be performed *before* reduction
517 to rule out fracture. No MRI is performed after the reduction, but a CT scan
518 may be performed in addition to the X-ray after reduction.

519 Discussion

520 The results of this study provide several insights into the potential use of AI chatbots
521 like ChatGPT in healthcare, particularly in enhancing patient education during the
522 surgical informed consent consultation procedure.

523 *What are patients' preferences regarding the use of an AI chatbot in conjunction with*
524 *a physician in surgical informed consent in comparison to a physician only?*

525 Study results showed a clear expressed need for a human doctor in the informed
526 consent procedure. No shifts in patients' expressed need were observed during the
527 study procedure. Results also show a preference for scenario A, a human doctor only
528 with worse information-delivery vs. scenario B, an AI-assistant only, with better
529 information-delivery. The preferred scenario (i.e., scenario A) did not change during
530 the procedure of informed consent. However, the preference strength in each group
531 differed at $t = 1$ from 0.83 in the ChatGPT+RAG group to 1.83 in the native ChatGPT
532 group with 1 = slight preference for scenario A and 2 = moderate preference for
533 scenario A. The data presented suggest, that patients would prefer the presence of a
534 physician during the informed consent procedure. However, their preference for a
535 human-only scenario (scenario A) over an AI chatbot-only scenario (scenario B) was
536 only slight to moderate. This finding proposes the potential and additional benefit of
537 patients' interaction with AI chatbots.

538 In our study, state and trait anxiety exhibited no explanatory power for patients'
539 attitudes toward the necessity of a human doctor in the informed consent
540 consultation, as well as their preference for a human doctor over an AI chatbot. In
541 contrast, general attitude towards AI (GAAIS) showed significant correlation with their
542 preference towards interaction with AI during informed consent. One would assume,
543 that patients experiencing higher levels of anxiety may perceive AI-driven informed
544 consent as less reliable compared to direct human interaction, potentially due to a

545 need for empathetic communication. While it may be due to a lack of power with 36
546 patients, further research is necessary.

547 Li et al. [39] explored the concerns surrounding the integration of AI into healthcare
548 from both patient and physician perspectives. Key concerns are "algorithm aversion",
549 robophobia, loss of humanistic care, and interaction challenges. Patients are
550 particularly concerned about trust, privacy, and AI's ability to deal with rare
551 conditions or individual circumstances. Physicians, on the other hand, are concerned
552 about job security, legal liability, and the emotional toll of introducing AI technologies
553 into their work environment. The authors emphasize the importance of fostering
554 trust, improving AI-human interaction, and providing appropriate education and
555 policy frameworks to mitigate these concerns.

556 Emotional states of patients, such as anxiety towards AI, should be specifically
557 addressed when implementing AI systems in sensitive areas like informed consent.

558 Providing additional support, such as hybrid models combining AI with physician
559 oversight, could help alleviate patient concerns and improve acceptance, particularly
560 in patients with higher levels of anxiety.

561 *How does patient satisfaction compare across informed consent procedures involving*
562 *only a physician, a physician and ChatGPT, or a physician and RAG-integrated*
563 *ChatGPT?*

564 After the Q&A sessions all groups are similarly satisfied with their respective level of
565 informedness and the quantity of information they received. This means that the
566 Q&A with the AI chatbot was able to satisfy the information needs of the patients
567 similarly well compared to the physician answering questions. In the ChatGPT+RAG
568 group the AI-assistant was able to close the previously existing gaps in both level of
569 informedness and the quantity of information they received after the FAQ session.

570 Process of information delivery - In the ChatGPT+RAG group, the patients were more
571 satisfied with the process of information delivery after the Q&A session compared to
572 the human only and native ChatGPT groups, which may be due to the quality of
573 responses observed with that approach (discussed below). The physicians involved in
574 the study also reported that especially the more concise responses in the
575 ChatGPT+RAG group was seen very positive by the patients.

576 The improved satisfaction in the RAG group may be attributed to the more context-
577 specific and evidence-based responses generated by the RAG-enhanced system.
578 These findings are consistent with prior research indicating that the RAG approach
579 improves output quality, increases patient confidence, and understanding during
580 medical consultations in patient-AI chatbot interaction [40-42]. By tailoring responses
581 to individual questions and incorporating subject- and clinic-specific data, the RAG
582 approach appears to address key patient concerns more effectively than the native
583 ChatGPT model.

584 One of the most striking findings was that the control group, which did not interact
585 with the AI chatbot, asked the fewest questions (20) whereas the ChatGPT group
586 asked more (39), and the RAG group asked the most questions (52). It could have
587 been possible that the responses provided in the intervention groups (native ChatGPT
588 and ChatGPT+RAG) were not good, so the patients had to ask follow-up questions.
589 However, study results show that the responses in the interventions groups were
590 rather good in the native ChatGPT group and very good in the ChatGPT+RAG group
591 (see next section). A general indication of these results is therefore that patients in
592 informed consent procedures can benefit from multimedia decision aids and tools
593 such as ChatGPT that may help to encourage engagement and improve patient
594 understanding [43].

595 *How does the integration of RAG with ChatGPT improve the quality of responses to*
596 *patients' questions compared to the native ChatGPT model?*

597 The results clearly demonstrate the advantages of the RAG approach over native
598 ChatGPT in all measured quality dimensions. Responses generated with
599 ChatGPT+RAG consistently scored higher in relevance, accuracy, and clarity,
600 completeness, evidence-based, and length compared to those generated by the
601 native ChatGPT model. This demonstrates the effectiveness of incorporating domain-
602 specific information into AI chatbot responses to ensure better alignment with
603 patient needs.

604 ChatGPT+RAG responses were significantly more concise than native ChatGPT
605 responses, which tended to be much longer and more detailed. While detailed
606 responses may seem advantageous, overly long responses run the risk of
607 overwhelming patients and diluting the focus on their specific questions. In contrast,
608 the concise responses of ChatGPT+RAG were likely perceived as clearer and more
609 directly relevant, contributing to the higher quality ratings observed.

610 The use of RAG significantly reduced hallucinations in our study. In four out of five
611 cases in which the AI chatbot hallucinated in the ChatGPT+RAG group, hallucinations
612 were due to missing information in the FAQ document provided as information to the
613 AI chatbot. This might suggest that when properly designed and managed, AI
614 chatbots can provide responses with very few hallucinations. Other studies confirmed
615 the effectiveness of a RAG approach für medical AI chatbots [40-42,44,45]. The low
616 error rate obtainable through RAG in conjunction with as complete as possible and
617 guideline-compliant FAQ documentation, together with the high consistency of an AI
618 chatbot, can therefore lead to very good results in combination with a human in
619 medical education. Furthermore, the minimal hallucinations observed in the

620 ChatGPT+RAG group provided valuable insights, enabling improvements to our
621 internal FAQ documentation following the study.

622 These results show the value of the RAG approach optimizing AI chatbots for use case
623 specific and clinic specific information, particularly in sensitive processes like surgical
624 informed consent consultations, where correct, clear, and comprehensive
625 communication is essential. However, it is important to continuously monitor these
626 tools to ensure patient safety in all cases. Also, more research is needed with
627 different datasets provided to the AI chatbot via the RAG approach, different AI-
628 language models, different RAG-architectures, different prompts, and with more
629 patients.

630 *How do patient sociodemographic characteristics and general attitude towards AI*
631 *influence patients' preference towards using an AI chatbot in surgical informed*
632 *consent?*

633 Our results show that age, gender, and education did not influence patients'
634 preference towards using an AI chatbot in surgical informed consent. There is no clear
635 consensus as per today in the body of literature about the role of age in AI ore
636 technology acceptance. Witkowski et al found that age is associated with AI
637 discomfort in healthcare [46]. Miller et al. who studied patients' utilization and
638 perception of an artificial intelligence-based symptom assessment and advice
639 technology in a British primary care waiting room identified a number of age-related
640 trends among respondents, with a directional trend for more young respondents (18-
641 24-year olds) to report that the tool provided helpful advice than older respondents
642 (adults aged 70+) [47]. Other studies show that older people are not per se "anti-
643 tech" or "anti-AI". Xiang et al. found that receptivity to medical AI increases with age,
644 reaching its highest levels among individuals aged 49 years and older [48]. More
645 research is required in different age groups and medical domains, e.g. in arthroscopy
646 where patients are usually younger.

647 Patients' general attitudes toward artificial intelligence demonstrated weak
648 explanatory power for patients' preference regarding the necessity of a doctor in the
649 informed consent process and their preference for a human doctor over an AI
650 chatbot. However, further research is needed to better understand the impact of
651 attitudes towards AI. The general attitude of patients towards AI could be considered
652 in the design of processes and systems in the future. This is also against the
653 background that AI and systems that integrate AI are developing rapidly and their
654 implementation in interaction with patients and doctors must therefore be
655 continuously examined.

656 The results of the present study demonstrate that, although the analysis of native
657 ChatGPT outputs shows significantly worse performance compared to ChatGPT+RAG,
658 patients report high satisfaction with the process of information delivery, their
659 perceived level of informedness, and the amount of information received across all
660 three groups. This should be considered in the design of AI driven patient
661 interactions. Despite the suboptimal quality of native ChatGPT responses as
662 evaluated by physicians, the RAG approach showed significantly better performance,
663 including higher patient engagement and markedly fewer hallucinations, making it a
664 more promising option for precise patient information delivery.

665 However, given that patients prefer the presence of a physician and that the RAG
666 approach cannot eliminate hallucinations, we strongly recommend using RAG-based
667 AI chatbots, but only in combination with a physician. Ensuring that patients are not
668 left alone in the information process is crucial. RAG chatbots, meanwhile, offer a
669 complementary solution for enhancing patient communication and reducing the time
670 burden on physicians in daily practice. The additional time gained can help address
671 issues related to patient education, literacy barriers, language differences, and limited
672 doctor communication skills (e.g., offering an easy-language mode).

673 *Future Research Directions*

674 The findings of this study suggest several potential directions for future research to
675 better understand patient preferences and concerns regarding AI chatbots in
676 healthcare, as well as to explore their possible applications in patient care.

677 Since the RAG approach has proven a significant higher quality compared to the
678 native ChatGPT version, it might be advisable to prioritize further investigation and to
679 expand the approach by incorporating patient-specific input. Further investigation is
680 needed to assess AI chatbots that could provide patient tailored information and
681 guidance during preoperative and postoperative care.

682 Future studies should also explore patient interactions using different interfaces (e.g.,
683 keyboards or voice assistants) and mobile applications of AI chatbots, particularly
684 outside formal informed consent procedures.

685 From a technological perspective, further work is needed to examine how different
686 implementations such as variations in the RAG framework or AI-language models and
687 variations in the type and structure of information provided to and by the AI chatbot.

688 The potential for these tools to be perceived as reliable and trustworthy by patients
689 requires careful evaluation while adhering to the principles of patient-centered care
690 [46].

691 Finally, demographic factors such as age, cultural background, and other
692 sociodemographic variables need to be studied in greater detail to understand how
693 different patient groups perceive and interact with AI chatbots. While AI has the
694 potential to make healthcare more accessible, it is essential to ensure that these tools
695 address the diverse needs of patients and do not inadvertently create new barriers to
696 care.

697 In our study we aimed at covering a wide range of questions regarding the new field
698 of patient AI chatbot interaction in the surgical informed consent procedure. Yet, due
699 to the exploratory nature is not without limitations. The over-representation of older
700 patients in our study may limit the applicability of the findings to a more diverse
701 patient population. In addition, the presence of a physician during all informed
702 consent consultations in all three groups likely influenced the dynamics of patient
703 interactions with the AI assistant, providing a setting that differed from direct,
704 independent patient-AI interactions. The involvement of two different physicians in
705 conducting the informed consent procedures may have introduced variability in
706 communication styles.

707 Before AI chatbots are widely implemented, future studies should aim to address
708 these limitations by including demographically diverse patient populations, examining
709 different degrees of indirect to more direct patient-AI interactions, and ensuring
710 consistency of the results by further studies with different practitioners and different
711 hospitals. These efforts will help validate our findings and provide a more
712 comprehensive understanding of the role AI chatbots can play in informed consent
713 procedures and patient care.

714 5. Conclusion

715 This study is the first exploring the innovative use of AI chatbots, specifically ChatGPT
716 and its RAG-enhanced version, in preoperative patient education during surgical
717 informed consent consultations. The results are promising, suggesting that AI
718 chatbots, particularly when customized with clinical and hospital-specific data via
719 RAG, can provide mostly accurate, context-specific, and concise information that
720 enhances patient satisfaction and engagement.

721 This study lays the groundwork for further research into how AI chatbots can be
722 integrated into diverse healthcare contexts. Future studies should focus on refining AI
723 chatbots for medical use, testing them in more autonomous patient interactions, and
724 exploring their impact across various demographics and clinical scenarios.

725 With thoughtful development and implementation, AI chatbots have the potential to
726 transform patient education, improve healthcare efficiency, and contribute to better
727 outcomes. However, as we embrace this opportunity, it is essential to remain cautious
728 about challenges such as trust, data privacy, and the preservation of the human touch
729 in patient care.

Acknowledgments

The authors thank Dr. Moses El Kayali for assistance in identifying appropriate data collection instruments.

Conflicts of Interest

None declared.

Abbreviations

AI: artificial intelligence

GAAIS: General Attitudes towards Artificial Intelligence Scale

GPT: generative pre-trained transformers

HLT: health literacy test

PPSQ: patients' preferences and satisfaction questionnaire

RAG: retrieval-augmented generation

SDQ: patient sociodemographic questionnaire and class-index

STAI: State-Trait-Anxiety Inventory

THA: total hip arthroplasty

References

- Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;388(1233):1233-1239. [doi:10.1056/NEJMSr2214184]
- Draschl A, Hauer G, Fischauer SF et al. Are ChatGPT's free-text responses on periprosthetic joint infections of the hip and knee reliable and useful? *J Clin Med* 2023;12(6655):6655. [doi:10.3390/jcm12206655]
- Li Y, Liang S, Zhu B et al. Feasibility and effectiveness of artificial intelligence-driven conversational agents in healthcare interventions: a systematic review of randomized controlled trials. *Int J Nurs Stud* 2023;143(104494):104494. [doi:10.1016/j.ijnurstu.2023.104494]
- Gabarron E, Larbi D, Denecke K, Årsand E. What do we know about the use of chatbots for public health? *Stud Health Technol Inform* 2020;270(796):796-800. [doi:10.3233/SHTI200270]
- Caldarini G, Jaf S, McGarry K. A literature survey of recent advances in chatbots. *Information* 2022;13(41):41. [doi:10.3390/info13010041]
- Shah A, Wahood S, Guermazi D, Brem CE, Saliba E. Skin and syntax: large language models in dermatopathology. *Dermatopathology* 2024;11(101):101-111. [doi:10.3390/dermatopathology11010009]
- 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 Intern Med* 2023;183(589):589-596. [doi:10.1001/jamainternmed.2023.1838]
- Suppadungsuk S, Thongprayoon C, Miao J et al. Exploring the potential of chatbots in critical care nephrology. *Medicines* 2023;10(58):58. [doi:10.3390/medicines10100058]
- Kienzle A, Niemann M, Meller S, Gwinner C. ChatGPT may offer an adequate substitute for informed consent to patients prior to total knee arthroplasty—yet caution is needed. *J Pers Med* 2024;14(69):69. [doi:10.3390/jpm14010069]
- Li J, Cheng X, Zhao X, Nie J-Y, Wen J-R. HaluEval: a large-scale hallucination evaluation benchmark for large language models. *Proc 2023 Conf Empir Methods Nat Lang Process* 2023;6449-6464. [doi:10.18653/v1/2023.emnlp-main]
- OpenAI, Achiam J, Adler S et al. GPT-4 technical report. *arXiv* 2023. [doi:10.48550/arXiv.2303.08774]
- Lee GU, Hong DY, Kim SY et al. Comparison of the problem-solving performance of ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard for the Korean emergency medicine board examination question bank. *Medicine* 2024;103(e37325):e37325. [doi:10.1097/MD.00000000000037325]
- Xue E, Bracken-Clarke D, Iannantuono GM, Choo-Wosoba H, Gulley JL, Floudas CS. Utility of large language models for health care professionals and patients in navigating

- hematopoietic stem cell transplantation: comparison of the performance of ChatGPT-3.5, ChatGPT-4, and Bard. *J Med Internet Res* 2024;26(e54758):e54758. [doi:10.2196/54758]
- Alessandri-Bonetti M, Liu HY, Donovan JM, Ziembicki JA, Egro FM. A comparative analysis of ChatGPT, ChatGPT-4, and Google Bard performances at the advanced burn life support exam. *J Burn Care Res* 2024;45(945):945-948. [doi:10.1093/jbcr/irae044]
- Sabaner MC, Hashas ASK, Mutibayraktaroglu KM, Yozgat Z, Klefter ON, Subhi Y. The performance of artificial intelligence-based large language models on ophthalmology-related questions in Swedish proficiency test for medicine: ChatGPT-4 omni vs Gemini 1.5 Pro. *AJO Int* 2024;1(100070):100070. [doi:10.1016/j.ajoint.2024.100070]
- Shahriar S, Lund B, Mannuru NR et al. Putting GPT-4o to the sword: a comprehensive evaluation of language, vision, speech, and multimodal proficiency. *arXiv* 2024. [doi:10.48550/arXiv.2407.09519]
- Goodman RS, Patrinely JR, Osterman T, Wheless L, Johnson DB. On the cusp: considering the impact of artificial intelligence language models in healthcare. *Med* 2023;4(139):139-140. [doi:10.1016/j.medj.2023.02.008]
- Deutsche Gesellschaft für Orthopädie und Unfallchirurgie e.V. (DGOU). Evidenz- und konsensbasierte Indikationskriterien zur Hüfttotalendoprothese bei Coxarthrose. S3-Leitlinie. 2021 Mar 24. Available from: https://register.awmf.org/assets/guidelines/187-0011_S3_Indikationskriterien_H%C3%BCfttotalendoprothese_bei_Coxarthrose_2021-04.pdf
- Wainwright TW, Gill M, McDonald DA et al. Consensus statement for perioperative care in total hip replacement and total knee replacement surgery: Enhanced Recovery After Surgery (ERAS®) Society recommendations. *Acta Orthop* 2019;91(1):1-17. [doi:10.1080/17453674.2019.1683790]
- Zampogna B, Parisi FR, Zampoli A, Papalia R. Artificial intelligence in orthopaedic surgery made easy. *Springer* 2024;43-53. [doi:10.1007/978-3-031-70310-2_5]
- Learmonth ID, Young C, Rorabeck C. The operation of the century: total hip replacement. *Lancet* 2007;370(1508):1508-1519. [doi:10.1016/S0140-6736(07)60457-7]
- Trousdale RT, McGrory BJ, Berry DJ, Becker MW, Harmsen WS. Patients' concerns prior to undergoing total hip and total knee arthroplasty. *Mayo Clin Proc* 1999;74(978):978-982. [doi:10.4065/74.10.978]
- Macario A, Schilling P, Rubio R, Bhalla A, Goodman S. What questions do patients undergoing lower extremity joint replacement surgery have? *BMC Health Serv Res* 2003;3(11):11. [doi:10.1186/1472-6963-3-11]
- Marinelli V, Danzi OP, Mazzi MA et al. PREPARE: preoperative anxiety reduction. One-year feasibility RCT on a brief psychological intervention for pancreatic cancer patients prior to major surgery. *Front Psychol* 2020;11(362):362. [doi:10.3389/fpsyg.2020.00362]
- Chiu PL, Li H, Yap KY, Lam KMC, Yip PLR, Wong CL. Virtual reality-based intervention to reduce preoperative anxiety in adults undergoing elective surgery. *JAMA Netw Open* 2023;6(e2340588):e2340588. [doi:10.1001/jamanetworkopen.2023.40588]

- 829 Sepucha KR, Atlas SJ, Chang Y et al. Informed, patient-centered decisions associated with
830 better health outcomes in orthopedics: prospective cohort study. *Med Decis Mak*
831 2018;38(1018):1018-1026. [doi:10.1177/0272989X18801308]
- 832 Narayanan AS, Stoll KE, Pratson LF, Lin FC, Olcott CW, Del Gaizo DJ. Musculoskeletal
833 health literacy is associated with outcome and satisfaction of total knee arthroplasty. *J*
834 *Arthroplast* 2021;36(S192):S192-S197. [doi:10.1016/j.arth.2021.02.075]
- 835 Schepman A, Rodway P. Initial validation of the general attitudes towards artificial
836 intelligence scale. *Comput Hum Behav Rep* 2020;1(100014):100014.
837 [doi:10.1016/j.chbr.2020.100014]
- 838 CD S, RL G, Lushene R. The manual for the State-Trait Anxiety Inventory. Consulting
839 Psychologists Press 1970.
- 840 Grimm J. State-Trait-Anxiety Inventory nach Spielberger. Deutsche Lang- und Kurzversion.
841 MF-Working Paper 2009.
- 842 Magruder ML, Rodriguez AN, Wong JCJ et al. Assessing ability for ChatGPT to answer total
843 knee arthroplasty-related questions. *J Arthroplast* 2024;39(2022):2022-2027.
844 [doi:10.1016/j.arth.2024.02.023]
- 845 Deutsche Gesellschaft für Orthopädie und Unfallchirurgie e.V. (DGOU). Evidenz- und
846 konsensbasierte Indikationskriterien zur Hüfttotalendoprothese bei Coxarthrose. S3-Leitlinie.
847 2021 Mar 24. Available from: [https://register.awmf.org/assets/guidelines/187-](https://register.awmf.org/assets/guidelines/187-0011_S3_Indikationskriterien_H%C3%BCfttotalendoprothese_bei_Coxarthrose_2021-04.pdf)
848 [0011_S3_Indikationskriterien_H%C3%BCfttotalendoprothese_bei_Coxarthrose_2021-04.pdf](https://register.awmf.org/assets/guidelines/187-0011_S3_Indikationskriterien_H%C3%BCfttotalendoprothese_bei_Coxarthrose_2021-04.pdf)
- 849 AE – Deutsche Gesellschaft für Endoprothetik e.V. Häufig gestellte Fragen zur Hüfte.
850 Available from: <https://www.ae-germany.com/patienten/haeufig-gestellte-fragen-zur-huefte>
- 851 Huang L, Yu W, Ma W et al. A survey on hallucination in large language models: principles,
852 taxonomy, challenges, and open questions. *ACM Trans Inf Syst* 2024; [doi:10.1145/3703155]
- 853 Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive
854 summarization. *Proc 58th Annu Meet Assoc Comput Linguistics* 2020;1906–1919.
855 [doi:10.18653/v1/2020.acl-main.173]
- 856 Rawte V, Chakraborty S, Pathak A et al. The troubling emergence of hallucination in large
857 language models - an extensive definition, quantification, and prescriptive remediations.
858 *arXiv* 2023. [doi:10.48550/arxiv.2310.04988]
- 859 Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharm Stat*
860 2025;4(287):287–291. [doi:10.1002/pst.185]
- 861 Landis JR, Koch GG. The measurement of observer agreement for categorical data.
862 *Biometrics* 1977;33:159. [doi:10.2307/2529310]
- 863 Li W, Liu X. Anxiety about artificial intelligence from patient and doctor-physician. *Patient*
864 *Educ Couns* 2025;133:108619. [doi:10.1016/j.pec.2024.108619]

- 865 Sanna L, Bellan P, Magnolini S et al. Building certified medical chatbots: overcoming
866 unstructured data limitations with modular RAG. Proceedings of the First Workshop on
867 Patient-Oriented Language Processing 2024;124–130.
- 868 Li A, Shrestha R, Jegatheeswaran T, Chan HO, Hong C, Joshi R. Mitigating hallucinations in
869 large language models: a comparative study of RAG-enhanced vs. human-generated medical
870 templates. MedRxiv 2024;2024.09.27.24314506. [doi:10.1101/2024.09.27.24314506]
- 871 Kirubakaran S, Kathrine JW, Kanaga EG et al. A RAG-based medical assistant especially for
872 infectious diseases. 2024 Int Conf Inven Comput Technol (ICICT) 2024;00:1128–1133.
873 [doi:10.1109/ICICT60155.2024.10544639]
- 874 Porter AL, Ebot J, Lane K et al. Enhancing the informed consent process using shared
875 decision making and consent refusal data from the CLEAR III trial. Neurocritical Care
876 2020;32:340–347. [doi:10.1007/s12028-019-00860-y]
- 877 Lewis P, Perez E, Piktus A et al. Retrieval-augmented generation for knowledge-intensive
878 NLP tasks. arXiv 2020. [doi:10.48550/arxiv.2005.11401]
- 879 Martino A, Iannelli M, Truong C. The Semantic Web: ESWC 2023 satellite events,
880 Hersonissos, Crete, Greece, May 28 - June 1, 2023, proceedings. Lect Notes Comput Sci
881 2023;182–185. [doi:10.1007/978-3-031-43458-7_34]
- 882 Witkowski K, Okhai R, Neely SR. Public perceptions of artificial intelligence in healthcare:
883 ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024;25:74.
884 [doi:10.1186/s12910-024-01066-4]
- 885 Miller S, Gilbert S, Virani V, Wicks P. Patients' utilization and perception of an artificial
886 intelligence-based symptom assessment and advice technology in a British primary care
887 waiting room: exploratory pilot study. JMIR Hum Factors 2020;7:e19713.
888 [doi:10.2196/19713]
- 889 Xiang Y, Zhao L, Liu Z et al. Implementation of artificial intelligence in medicine: status
890 analysis and development suggestions. Artif Intell Med 2020;102:101780.
891 [doi:10.1016/j.artmed.2019.101780]