Original Article

Use of an Artificial Intelligence Conversational Agent (Chatbot) for Hip Arthroscopy Patients Following Surgery

Tim Dwyer, M.B.B.S., Ph.D., F.R.A.C.S., F.R.C.S.C., Graeme Hoit, M.D., David Burns, M.D., Ph.D., F.R.C.S.C., James Higgins, M.D., F.R.C.S.C., Justin Chang, M.D., F.R.C.S.C., Daniel Whelan, M.D., M.Sc., F.R.C.S.C., Irene Kiroplis, and Jaskarndip Chahal, M.D., M.Sc., M.B.A., F.R.C.S.C.

Purpose: The purpose of this study was to evaluate the use of an AI conversational agent during the postoperative recovery of patients undergoing elective hip arthroscopy. **Methods:** Patients undergoing hip arthroscopy were enrolled in a prospective cohort for their first 6 weeks following surgery. Patients used standard SMS text messaging to interact with an artificial intelligence (AI) chatbot ("Felix") used to initiate automated conversations regarding elements of postoperative recovery. Patient satisfaction was measured at 6 weeks after surgery using a Likert scale survey. Accuracy was determined by measuring the appropriateness of chatbot responses, topic recognition, and examples of confusion. Safety was measured by evaluating the chatbot's responses to any questions with potential medical urgency. **Results:** Twenty-six patients were enrolled with a mean age of 36 years, and 58% ($n = 15$) were male. Overall, 80% of patients ($n = 20$) rated the helpfulness of Felix as good or excellent. In the postoperative period, 12/25 (48%) patients reported being worried about a complication but were reassured by Felix and, thus, did not seek medical attention. Of a total of 128 independent patient questions, Felix handled 101/128 questions appropriately (79%), either by addressing them independently, or facilitating contact with the care team. Felix was able to adequately answer the patient question independently 31% of the time ($n = 40/128$). Of 10 patient questions that were thought to potentially represent patient complications, in 3 cases Felix did not adequately address or recognize the health concern—none of these situations resulted in patient harm. **Conclusion:** The results of this study demonstrate that the use of a chatbot or conversational agent can enhance the postoperative experience for hip arthroscopy patients, as demonstrated by high levels of patient satisfaction. **Levels of Evidence:** Level IV, therapeutic case series.

Introduction

Innovation in healthcare technology has allowed the possibility of augmenting the patient experience. The use of technology such as automated text messages has been used in patients following hip arthroscopy, with high levels of patient satisfaction.[1] There have also been studies focused on reducing in-person follow-up with the use of care delivered by mobile applications, allowing the uploading of pictures from a mobile smartphone for wound checks, as well as the ability of clinicians to respond to clinical questions. These remote methods of follow-up have been shown to be safe, convenient, and have high satisfaction from orthopaedic surgical patients following anterior cruciate ligament reconstruction,[2] knee arthroplasty,[3] and spine surgery.[4] However, these applications (i.e., "apps") still require review and response by surgeons and their staff in a timely manner.

Conversational agents (chatbots) are computer programs designed to simulate conversation with human users. With modern advances in natural language processing techniques, artificial intelligence (AI), and computing power, conversational agents have been developed for diverse applications in medicine, including screening for health conditions, counseling, and at-home health management support.[5]. Although

From the Women's College Hospital, Toronto, Canada (T.D., G.H., D.W., J.C.); Mount Sinai Hospital, Toronto, Canada (T.D.); and University of Toronto Orthopedic Sports Medicine, Toronto, Canada (T.D., G.H., D.B., J.H., J.C., D.W., I.K.)

Received October 23, 2022; accepted January 31, 2023.

Address correspondence to Women's College Hospital, Toronto M5S1B2, Canada. E-mail: tim.dwyer@wchospital.ca

2666-061X/221280

https://doi.org/10.1016/j.asmr.2023.01.020

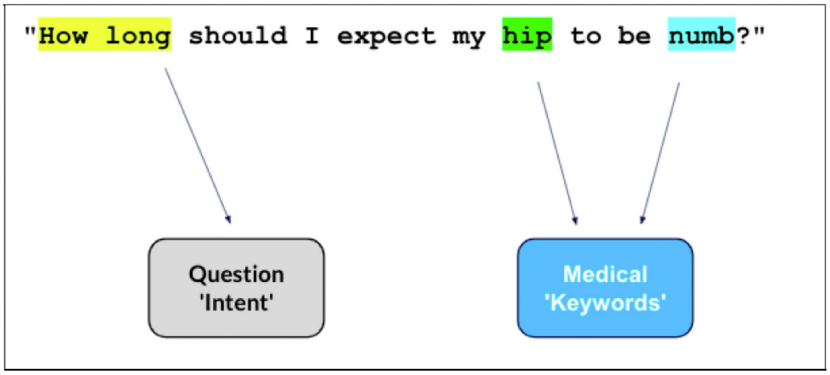
the use of chatbots in healthcare is a new and evolving field, acceptability and patient experience have so far been largely positive.[6]

The utilization of AI conversational agents to provide medication reminders, rehabilitation instructions, and answer patients' questions and concerns can potentially enhance quality of care, while reducing the burden on health resources. At this time, studies have investigated the use of chatbots in orthopedics to guide and reduce opioid use after trauma surgery[7] and provide health education following surgery.[8]

The purpose of this study was to evaluate the use of an AI conversational agent during the postoperative recovery of patients undergoing elective hip arthroscopy. We hypothesized that an AI chatbot could accurately and safely monitor patients following hip arthroscopy in a manner that was satisfactory to patients.

Methods

This study was conducted with ambulatory care patients undergoing hip arthroscopy, conducted over the first 6 weeks of care following surgery. Inclusion criteria were 1) patients undergoing hip arthroscopy for femoroacetabular impingement (FAI) syndrome, 2) patients between the ages of 18 and 60, 3) patients able to use a mobile device with an SMS plan and a camera, and 4) patients able to communicate and read in English. Exclusion criteria were 1) patients unable to take narcotic pain medication postoperatively and 2) patients ongoing litigation or compensation claims secondary to hip problems.



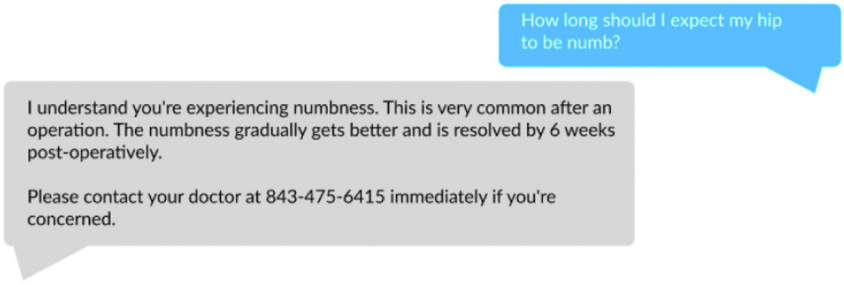


Fig 1. Example of natural language processing to understand patients' questions and concerns via SMS.

Fig 2. Examples of primary categories of responses to patients.

All patients consented to undergo hip arthroscopy after a minimum of 3 months of nonoperative treatment. Patients were identified as having FAI syndrome based upon history, clinical examination, radiograph, and magnetic resonance imaging. All patients had a general anesthetic and local nerve block, and surgery was performed on a traction table using a perineal post. Patients underwent hip arthroscopy with labral repair or debridement, and a pincer, Cam, or combined bony resection, as clinically indicated by one of three fellowship-trained hip arthroscopy surgeons. Patients underwent a standardized postoperative rehabilitation protocol and were placed touch weight bearing on crutches for the first 6 weeks. Patients continued to attend standard clinic follow-up at 2 and 6 weeks postoperatively.

Artificial Intelligence System and Automated Conversation

Patients interacted with an AI-powered Virtual Health Coach, nicknamed "Felix", throughout their recovery using standard SMS text messaging (Memora Health Inc., San Francisco, CA). Three fellowship-trained hip arthroscopy surgeons provided written protocols, handouts, and postoperative discharge instructions for content creation. Following this, Felix was customized for hip arthroscopy by creating an internal repository / care program of specific content divided into three different categories: instructional tips and reminders, surveys of patient data, and automatic question support.

Following the creation of a care program, Felix was used to process and maintain automated conversations with patients via SMS. The natural language-processing techniques involved are proprietary. However broadly speaking, three different mechanisms were used to determine what message or response was provided to a patient: context (the context of a conversation through the examination of prior messages); keywords (the medical and nonmedical keywords in each phrase); and intent (the intent of the phrase provided by the patient). Following this, hundreds of thousands of combinations of "intents" and "keywords" are mapped to single responses (Fig 1). Felix was able to respond to patients in several ways—provide automated guidance, ask follow-up questions, or forward a message to the care team (Fig 2).

Felix was programmed to provide automated pain management reminders, wound management guidance, physiotherapy, and rehabilitation instructions specific to hip arthroscopy, as well as perform general wellness checks (Table 1). Medication reminders were varied depending on surgeon preference, i.e., one surgeon used postoperative aspirin for DVT prophylaxis, while others did not. No proactive medical advice was given by Felix, such as switching or starting new medications. Surgeons were alerted to any concerns or "red flags" that Felix encounters via email.

Table 1. Automated Guidance

Category	Automated Guidance
Pain management	Good evening Robert. Pain is a normal part of the healing process. Taking your prescribed narcotic medication will help manage this. Some common side effects include nausea, drowsiness, and constipation. To decrease these side effects, take your medication with food. Only take as directed on the bottle.
	Pain medication was injected into your surgical site and will wear off within 8-12 hours.
	If constipation occurs, consider an over-the-counter laxative medication. If you are having problems with nausea and vomiting with your narcotic medication, please call the surgeon's office, as they may be able to change you to a different narcotic.
	This is a reminder to take your Celebrex 200 mg this evening. Have a good night!
	Good morning Robert! This is a reminder to take your Celebrex 200 mg this morning. You should be taking this medication twice daily (12 hours apart). This will help smooth out any pain "peaks and valleys" in between narcotic medication dosages.
	What is your current pain level on a scale of 0-10, where 0 is "no pain" and 10 is "worst pain you've ever felt"?
	Just a reminder that the recovery process is different for everyone. Some patients are pain-free in 3-6 months, and some even sooner. The key components of any successful recovery after surgery include rest, appropriate physical therapy, and time.
	Thank you for letting me know! Around 2 weeks, most patients begin to taper or fully stop the use of narcotics. These are addictive medications, so you should try to take the minimum possible to manage your pain.
Wound management	It is normal to experience some swelling and bleeding following surgery. If blood soaks through your dressing, reinforce it with an additional bandages. Your incision will be checked at your next follow-up appointment.
	For the first 2 days, all dressings should remain intact. Sponge bath for bathing.
	You may remove the surgical dressing on the 3rd day after surgery, which is today! Apply a clean bandage over your wound. Showering is fine as long as your wounds remain sealed with bandages. You should change the dressing if it becomes soaked with water.
	To avoid infection there should be no immersion (bath, pool, hot tub, lake, ocean, etc.) of the operative site for the first 3 weeks postoperatively. After 3 weeks, you will be allowed to do so!
	It is abnormal to experience significant drainage from your wounds 1 week after your surgery. Please call your surgeon's office if this occurs.
	Hi Robert! You're 6 weeks out, I wanted to check in. Do you have any concerns about your surgical site?
Physiotherapy and rehabilitation	Physiotherapy will begin 6 weeks after your surgery. Follow your surgeon's exercise protocol and the images I send you until your first appointment. Go to the physiotherapy clinic that is most convenient for you, usually one that is close to your work or home. They will follow your surgeon's rehabilitation plan.
	This is a reminder to make sure you are performing your 2 exercises 3 times a day (morning, lunch, and dinner):
	This is a reminder to complete your 2 exercises 3 times each day. If these have been difficult for you, it is important for you to remember that staying active will help you recover and prepare you to start physical therapy. Here are some pictures of the exercises if you need to refresh your memory
	A good recovery depends on staying active, so please try to do these exercises each day. If you find you are unable to do an exercise that I have sent you, you should not push yourself. It is still highly important that you are active while you recover and work with your therapist to try to improve your motion. Your surgeon will assess your motion at your next visit.
Emergencies and complications	If you have a persistent fever (over 101° Fahrenheit) or chills, call your surgeon's office. It is normal to have a low-grade fever for the first day or two following surgery. Take Tylenol to help reduce your fever.
	If you have unrelenting pain that is not controlled by your narcotic pain medication, or you have a sudden increase in pain that is out of proportion to how you have been feeling, please call your surgeon's office
	If you have any difficulty breathing, please proceed to your nearest emergency room.
	Numbness in the genitals/groin: This can occur from the traction applied during surgery. It is a well-known risk of surgery and will be relayed to your physician. This numbness typically takes a few weeks to resolve and will be monitored by your physician.

AI Customization

Surgeons initially created a list of 73 questions that hip arthroscopy patients might ask, with corresponding answers. A trial period was undertaken for more than a month on user testers hired from Amazon Mechanical Turk (Amazon, Seattle, WA), who asked a total of 573 questions—following this, a further 65 questions and answers were added. A month prior to the study date,

Table 2. Messages and Interactions Between Patients and Felix

Individual Messages	
Source	No. (%)
Felix total	2505 (72.4%)
Pre-programmed	2217 / 2505 (88.5%)
AI Automated Response	288 / 2505 (11.5%)
Patient	871 (25.2%)
Clinician	83 (2.4%)
Total	3459
Interactions	
Type	No. (%)
Introductory message	26 (1.6%)
Felix initiated survey question	396 (25.1%)
Medications	240 / 396 (60.6%)
Physical therapy	156 / 396 (39.4%)
Felix initiated text question	68 (4.3%)
Wound check	68 / 68 (100%)
Felix tip or reminder	987 (62.6%)
Administrative	8
Recovery	23
Medications	598
Preoperation instructions	80
Physical therapy	172
Wound care	106
Patient initiated questions	99 (6.3%)
Total	1576 (100%)

the chatbot was tested and further refined with the hip arthroscopy surgeons. This resulted in increased support postsurgery, messages shortened to increase readability, survey questions increasingly spread out to ensure patients weren't receiving multiple questions in one day, and exercise reminders alternated with exercises survey to boost engagement. Over the course of the study, following monthly review, another 32 questions and answers were added.

Felix's responses were also assessed by the surgeon group, research group, and Memora weekly throughout the study period. Questions that Felix was unable to answer correctly were fed back into the database with correct responses in order to improve the capability of the system.

Outcomes

The chatbot was evaluated in four domains: accuracy of responses to patient questions, safety of recommendations and actions, health resource utilization, and patient satisfaction.

Accuracy of Responses and Learning

Messages between patients and Felix were reviewed and organized into discrete interactions. Interactions containing patients' questions were analyzed qualitatively for accuracy of the AI-generated responses. For each interaction, we recorded the following: 1) appropriateness (whether Felix's action was appropriate), 2) topic recognition (whether or not Felix recognized the topic of the question [e.g., physical therapy, pain, wound issues etc.]), 3) confusion (whether or not Felix was confused at any point during the interaction), 4) self-awareness (whether or not Felix identified the confusion by stating so or asking clarifying questions), and 5) question management (how Felix dealt with the question (e.g., by answering it, by forwarding it to the surgeon, etc.). The number of times the surgeons had to correct or clarify Felix's responses was recorded. Felix's learning throughout the study period was assessed by comparing Felix's appropriateness, topic recognition, confusion, self-awareness, and question management of each interaction with the date of interaction initiation as a continuous variable.

Safety

Patient-initiated questions were reviewed by two reviewers, and any questions with medical urgency or where an incorrect response from Felix had the potential to cause harm were identified. Felix's responses to these questions were then analyzed for safety. Felix providing correct advice and/or directly forwarding the question to the surgeon or instructing the patient to seek immediate care were identified as "safe" responses. Incorrect advice and/or not contacting the surgeon or giving instruction to seek care were considered "unsafe" responses. Partially correct advice or action, such as asking follow-up questions without escalation or chatbot confusion was considered "potentially unsafe" responses.

Health Resource Utilization

The total number of health care telephone calls, emails, and health care visits (to surgeon, family physician and emergency department) related to surgery were captured at each postoperative visit by a study coordinator. Additionally, the number of times the patient messages to Felix were forwarded to the surgeon was recorded. These instances were reviewed by a member of the study team and evaluated for appropriateness (appropriate additional contact versus inappropriate additional contact). Additionally, we recorded the number of times Felix instructed patients to seek care without directly forwarding the message to the surgeon.

Patient Satisfaction

Patient adherence and response rate to Felix questions were recorded and analyzed. Patient satisfaction was measured at 6 weeks after surgery using a post-study patient survey. Satisfaction was measured on a

Table 3. Felix Answering Patient Question

Topics	
Patient Intended Topic of Question	No. (%)
Wound care	23 (18.0%)
Medication	17 (13.2%)
Physical therapy	16 (12.5%)
Restrictions	12 (9.4%)
Numbness	12 (9.4%)
Pain	10 (7.8%)
Administrative	9 (7.0%)
Sleeping	6 (4.7%)
Swelling/redness not at wound	5 (3.9%)
General recovery	5 (3.9%)
Constipation	2 (1.6%)
Mood	2 (1.6%)
Diet	2 (1.6%)
Radiograph	1 (0.8%)
Vomiting	1 (0.8%)
Weakness	1 (0.8%)
Preop preparation	1 (0.8%)
Multiple topics	1 (0.8%)
Emergency testing	1 (0.8%)
Introductory	1 (0.8%)
Total	128
Felix Understanding Questions	
Felix Recognized Topic of Interaction?	
Yes	74 / 128 (57.8%)
No	54 / 128 (42.2%)
Felix Confused During Interaction?	
Yes	80 / 128 (62.5%)
No	48 / 128 (37.5%)
Felix Recognized Confusion?	
Yes	54 / 80 (67.5%)
No	26 / 80 (32.5%)
Felix Handling of Questions	
Question Adequately Addressed by Felix Alone?	
Yes	40 / 128 (31.2%)
No	88 / 128 (68.8%)
Question Unaddressed by Felix, handled appropriately	
Forwarded to surgeon appropriately	36 / 88 (39.8%)
Not forwarded, told to seek care	17 / 88 (19.3%)
Patient did not answer follow-up/clarification questions	8 / 88 (10.2%)
Total	61 / 88 (69.3%)
Question Unaddressed by Felix, handled inappropriately	
Not contacted or told to seek care, physician override made	17 / 88 (17.0%)
No action taken	8 / 88 (11.4%)
Forwarded to surgeon inappropriately	2 / 88 (2.3%)
Total	27 / 88 (30.7%)

5-point Likert scale: excellent, good, fair, poor, and very poor, or strongly agree, agree, neutral, disagree, and strongly disagree, where appropriate.

Surgeon Satisfaction

Surgeon satisfaction was evaluated after completion of the study. Surgeons completed a questionnaire using the above 5-point Likert scales.

Statistical Analysis

Summary statistics of patient characteristics, messages and interactions, satisfaction survey responses, and patient-reported outcome measures (PROMs) were calculated using means, medians, ranges, counts, and percentages, where appropriate. Differences between preop and postop PROMs were compared using paired *t*-tests. Learning was assessed by comparing the binary coded variables of interaction appropriateness, topic recognition, confusion, self-awareness, and question management with the date of interaction initiation using univariate binary logistic regression. All tests for significance were two-sided and required a P value of $<.05$. All analyses were performed using SAS v9.4 (Cary, NC).

In calculating sample size for this pilot study, our goal was to identify problems in the chatbot responses with regard to safety, appropriateness, topic recognition, and confusion. Given the lack of published information at the time of study design with regard to the frequency of these problems in healthcare chatbots, we chose to calculate our sample, such that we could detect any issues with an incidence greater than 10% at 95% confidence. Using a previously described pilot study formula, a sample size of 25 patients was determined.[9]

Results

Between September 2019 and March 2020, 26 patients (58% male) were enrolled with a mean age of 26 (±9.4). All patients completed a 6-week follow-up.

A total of 3,459 messages were exchanged between patients and Felix over the study period—of these, 1,477 (94%) interactions were automated messages initiated by Felix (Table 2). There were 99 patient-initiated interactions, with a further 29 patient questions stemming from Felix-initiated interactions for a total of 128 patient questions (4.9 questions per patient, range: 0-33). The mean number of messages per interaction was 2.2 (range: 1-10).

Accuracy of Responses and Learning

A total of 128 patient questions was included in our analysis of AI-generated responses and learning (Table 3). In total, Felix handled 101 of the 128 questions appropriately (79%), either by addressing them independently, or facilitating contact with the care team (Fig 3). Felix was not able to handle 27/128 (21%) of questions appropriately. Of these, 17 were forwarded to the surgeon's office, 2 were forwarded to the surgeon's office inappropriately, and in a further 8 situations, it was deemed that no further action was necessary (Table 3). The likelihood of Felix adequately addressing the question did not change significantly over the study period ($P = .17$) (Table 4).

Felix was able to correctly identify the topic of the interaction 58% of the time ($n = 74/128$). Felix's ability

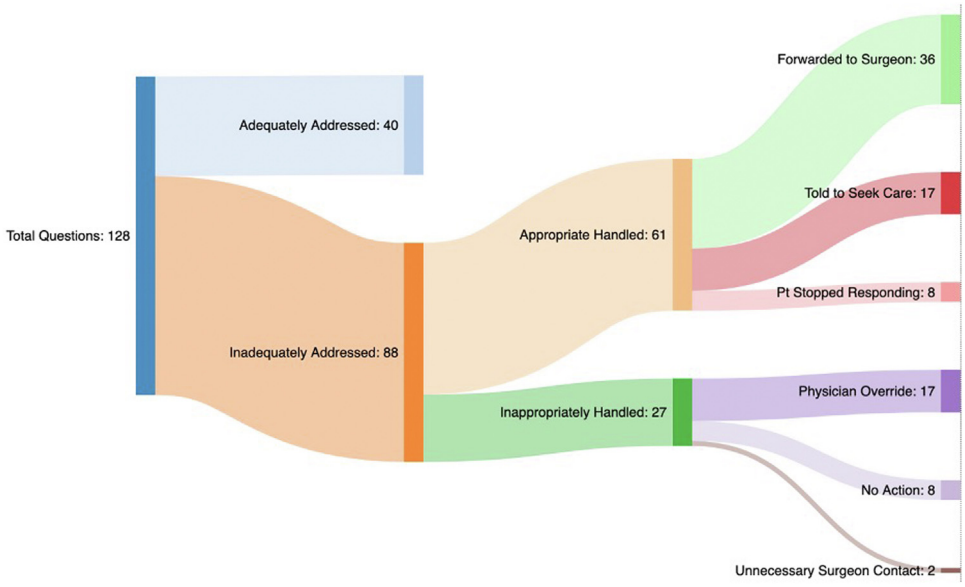


Fig 3. Flow diagram of Felix's responses to patient questions. In total, "Felix" handled 101 (40 adequately addressed, 61 appropriately handled) of the 128 questions appropriately (79%), either by adequately addressing them independently (40), or by appropriately handling questions and facilitating contact with the care team (61).

to identify the correct topic did not improve significantly over the study duration ($P = .42$). Felix was able to adequately answer the patient question independently 31% of the time ($n = 40/128$). This did improve significantly over the study period ($P = .040$).

Felix demonstrated confusion in 80 of the 128 interactions (63%), with examples given in Table 5. Confusion significantly reduced over time during the study period ($P = .033$).

Of the 74 instances in which Felix was confused, Felix recognized its confusion in 65% of these interactions ($n = 48/74$), termed self-awareness. Self-awareness did not improve significantly over the study duration ($P = .49$).

Safety

Of the 128 patient questions, 10 were identified as having safety implications, either because of medical urgency or because reviewers felt an incorrect response could result in harm. Of the 10 questions, Felix correctly responded to 7 of them (70%), either by providing correct advice, forwarding the question to the surgeon or instructing the patient to seek immediate care.

Three responses by Felix were identified as potentially unsafe by manual review. In the first, a patient described increased pain to the surgical site after 2 weeks, and Felix correctly clarified that the patient was talking about their incision but failed to follow-up on the symptoms further to rule out wound issues or relay the information to the care team. In the second, a patient described having pain and tightness in their calf after surgery. Here, Felix correctly responded, asking the patient to rate the severity of their pain, but after the patient failed to respond to this question, no further questions or escalating action was taken by Felix to rule out a deep vein thrombosis. In the third potentially unsafe response, the patient indicated that they thought

Table 4. Statistical Assessment of Felix Improvement Over Time

Felix Evaluation Category	Classification	Mean Age of Felix, days (95%CI)	Mean Difference, days (95% CI)	Odds Ratio, per 10 days (95% CI)	*P* Value*
Appropriateness ($n = 128$)	Appropriate action ($n = 101$)	104.2 (93.7-114.6)	16.8 (−7.2-50.6)	1.06 (0.98-1.14)	.17
	Inappropriate action ($n = 27$)	87.5 (61.7-113.3)		Ref	
Topic recognition ($n = 128$)	Recognized ($n = 74$)	104.1 (91.5-116.6)	8.1 (−11.7-28.0)	1.03 (0.96-1.09)	.42
	Not recognized ($n = 54$)	95.9 (80.0-111.9)		Ref	
Confusion ($n = 128$)	Not confused ($n = 48$)	112.8 (97.0-128.5)	19.4 (0.44-39.5)	1.07 (1.00-1.14)	.033
	Confused ($n = 80$)	93.4 (80.9-105.8)		Ref	
Self-awareness ($n = 80$)	Recognized confusion ($n = 54$)	96.4 (82.1-110.6)	9.3 (−17.4-36.0)	1.03 (0.95-1.12)	.49
	Did not recognize confusion ($n = 26$)	87.1 (61.2-112.9)		Ref	
Question management ($n = 128$)	Adequately addressed	114.7 (99.0-130.4)	20.7 (0.9-20.7)	1.07 (1.00-1.15)	.040
	Inadequately addressed	94.0 (81.7-106.3)		Ref	

95% CI = 95% confidence interval. n = number of patients; Ref, reference category.
*P values generated by logistic regression.

Table 5. Examples of Interactions and Classifications

Questions	Patient	Percentage
"Felix" confused but self-aware 54/79 (68%)	*PATIENT: "Will the bandages be removed at the follow-up appointment?"*	*FELIX: "Hmm.. I'm not sure what you're saying. Sorry! Try saying that in a different way. If this is an emergency, please call 911."*
Felix confused but self-unaware 25/79 (32%)	*PATIENT: "When can I put weight on my foot?"*	*FELIX: "I understand you're experiencing numbness. This is very common after an operation. The numbness gradually gets better and is resolved by 6 weeks postoperatively. Please contact your doctor at x immediately if you are concerned."*
Felix correctly addressed a question independently 36/127 (28%)	*PATIENT: "I feel capable of walking without crutches. Would this be bad for my recovery?"*	*FELIX: "You can put full weight on your operated limb 6 weeks after your surgery. Until then, your foot can be placed on the ground to help balance, but your weight shouldn't be placed through it."*
Felix appropriately reached out to surgeons, or instructed patients to do so 48/127 (38%)	*PATIENT: "There is a white string coming out of my incision. Is this normal? Is it a stitch?"*	*FELIX: "I'll forward that message to your doctor. If this is an emergency, please call 911."*
Felix was unable to adequately answer questions and did not inform surgeons 20/127 (16%)	*FELIX: "Do you have any concerns about your surgical site?"*	*PATIENT: "None. I'm wondering how and when the stitches come out. Thank you."* *FELIX: "Thanks for letting me know, I hope your recovery is going well."*
Felix facilitated unnecessary contact between patient and surgeon 2/127 (2%)	*FELIX: "Most patients are encouraged to taper their narcotic medication after 2 weeks. At this point, you should take as little narcotic medication as you can tolerate. If you are having trouble stopping, please call your care team at x.*	*PATIENT: "I haven't needed them in a while."* *FELIX: "I'll forward that message to your doctor. If this is an emergency, please call 911."*
Care team was required to intervene without the conversation being flagged by Felix 21/127 (17%)	*PATIENT: "I believe I have an infection in my surgical site. Do I go to hospital?"*	*FELIX: "I might be able to find that information for you if you rephrase your message."* *PATIENT: "My surgical site is hot, red and very painful."* *FELIX: "I still want to try and get you an answer, but if this is more urgent try calling your doctor."* *CLINICIAN OVERRIDE: "How is your hip feeling? Has the redness gone down? You can take a picture and send to Felix if you are worried about infection."*

they had a surgical site infection, and Felix was unable to recognize the intent of the patient's question. Felix failed to recognize the topic after the patient reworded their question twice until the clinical team stepped in to override and ask further questions. In each of these cases, no actual complication was identified at final follow-up. Ultimately, no harm occurred to patients secondary to Felix, and the oversight mechanisms in place during the study were also successful in mitigating the potential for harm.

Health Resource Utilization

Across the 26 participants, 13 patients combined for a total of 40 extra contacts with the healthcare system beyond the scheduled regular follow-up appointments over the first 6 postoperative weeks (Table 6). This

Table 6. Felix Encounters

Question	Yes	% Yes
Did you want to be in contact with your surgeon in the first 6 weeks?	7	28%
Was "Felix" sufficient in providing information needed before follow-ups within 6 weeks?	16	64%
Did you call the surgeon's office within the first 6 weeks?	4	16%
Did you email the surgeon's office within the first 6 weeks?	11	44%
Did you have an unscheduled health visit in the first 6 weeks?	3	12%
Extra contact (any call, email, or unscheduled visit)	13	52%
Did Felix alert you to a possible complication for which you did not seek medical attention?	1	4%
Were you worried about a complication but reassured by Felix, leading you to not seek medical attention?	12	48%
Did your surgeon alert you to a complication that was not identified or incorrectly identified by Felix?	0	0%

Table 7. Satisfaction Survey Results

Satisfaction Question Set 1 $n = 25$	Mean Score (SD)	Median Score (IQR)	No. of patients "Good" or "Excellent" (%)
Answers are on a scale of 1-5: (1=very poor, 2=poor, 3=fair, 4=good, and 5=excellent).			
My **overall satisfaction** with postoperative care after surgery was:	4.2 ± 0.9	4 (4-5)	20 (80%)
Overall, the **use of** the SMS messaging system, "**Felix**", that was involved in my postoperative recovery was:	4.0 ± 0.7	4 (4-4)	21 (84%)
Please rate the **helpfulness** of the SMS messaging system, Felix, as:	4.0 ± 0.8	4 (4-4)	20 (80%)
Please rate the **ease of use** of the SMS messaging system, Felix, as:	4.3 ± 0.7	4 (4-5)	22 (88%)
Please rate the **ease of understanding** of the SMS messaging system, Felix, as:	4.3 ± 0.7	4 (4-5)	22 (88%)
Please rate the **frequency of messages** you received from the SMS messaging system, Felix, as:	4.2 ± 0.8	4 (4-5)	21 (84%)
Please rate the **quality of messages** from the SMS messaging system, Felix, as:	3.9 ± 0.8	4 (3-4)	18 (72%)
Please rate the **likelihood of continuing** to use the SMS messaging system, Felix, in further postoperative scenarios as:	4.2 ± 0.8	4 (4-5)	21 (84%)
Please rate the **likelihood to recommend** this SMS messaging system, Felix, to others undergoing similar procedures as:	4.2 ± 1.0	5 (4-5)	19 (76%)
Please rate the **availability of your physicians** (i.e., how easy it was to get in contact with your doctor(s) when needed):	4.2 ± 0.8	4 (4-5)	22 (88%)
Satisfaction Questions Set 2 $n = 25$	Mean Score (SD)	Median Score (IQR)	No. patients "Agree" or "Strongly agree" (%)
Answers are on a scale of 1-5: (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, and 5=strongly agree).			
I **knew who to contact** if I had problems following discharge that could not be addressed by Felix.	4.5 ± 0.8	5 (4-5)	23 (92%)
Felix helped me understand what I needed to know in order to manage my condition at home.	4.1 ± 0.7	4 (4-4.5)	20 (80%)
Felix was able to provide information about how to **monitor my condition** for problems and danger signals.	4.1 ± 0.8	4 (4-5)	20 (80%)
My emotional needs (worries, fears, and anxieties) were acknowledged and addressed by Felix.	3.3 ± 1.1	3 (3-4)	17 (68%)
Felix was able to **diminish my worries**, fears, and anxieties surrounding my postoperative recovery.	3.6 ± 0.9	4 (3-4)	15 (60%)

n, number of patients; IQR, interquartile range; no., number; SD, standard deviation.

included 32 emails to the surgeon's office, 5 phone calls, and 3 emergency room/walk in clinic visits. Of the 32 emails, 11 were from 1atient. Twelve patients reported being worried about a complication but were reassured by Felix and, thus, did not seek medical attention. These instances were reviewed individually and were found to be appropriate.

Patient Satisfaction

Patients reported high levels of satisfaction with Felix's involvement in the postoperative care (mean 4.0/5 + 0.7) (Table 7). 80% of patients agreed or strongly agreed with the statement: "Felix helped me understand what I needed to know in order to manage my condition at home."

Surgeon Satisfaction

Overall, the three surgeons rated their overall satisfaction with Felix as excellent or good. The physician user portal was described as excellent by all three surgeons, and all three agreed or strongly agreed that they would be happy to continue to use Felix in further postoperative scenarios, and they would recommend Felix to other surgeons performing similar procedures.

Discussion

The results of this study demonstrate that the use of a chatbot or conversational agent can enhance the postoperative experience for hip arthroscopy patients, as demonstrated by high levels of patient satisfaction. However, while the accuracy of the chatbot's management of patients' questions reached above the threshold of 75%, there were many instances of chatbot confusion during the patient interaction that could be enhanced by further development. The improvement of the chatbot that occurred over time through the course of this study suggests that further refinement is achievable as the corpus of available relevant questions and responses is increased.

Previous studies in orthopedics have enhanced the postoperative period using automated text messages,[1] or with the use of mobile applications allowing patients to upload clinical photos and questions.[2] In contrast, a conversational agent (chatbot) is defined as a "computer program designed to stimulate conversation with human users".[6] Chatbots vary in complexity from simple models that can only respond if they recognized a keyword or a topic, to those with sophisticated natural language processing capabilities.[10] The use of chatbots in health care is increasing as a result of financial pressures, clinical demand, and a desire to increase efficiency in healthcare.[6] Elective ambulatory patients, such as hip arthroscopy patients, provide a very suitable patient population for this technology—a generally young patient age and a population that is typically adept and comfortable with mobile technologies. The most common topics of patient questions were in regard wound care (18%), medications (13%), physical therapy (13%), postoperative restrictions (9%), and numbness (9%). Each of these represent common and important concerns for patients' post hip arthroscopy—questions that would normally be directed to the surgeon's office. Importantly, in this study, 48% of patients reported being concerned about a complication and were appropriately reassured by Felix.

Systematic reviews of the use of chatbots in health care have reported a largely positive patient experience, as measured by patient satisfaction with chatbot interactions and responses.[6,11] Breast cancer patients have reported that chatbots can provide support to help follow their treatment plans,[12] while a study of the use of a chatbot with cancer patients demonstrated that engagement with the chatbot helped to reduce anxiety.[13] However, levels of engagement have been shown to drop off over time, with the former study demonstrating an initial 100% engagement rate that was reduced to 31% over an 8-month period.[12] At this time, the use of conversational agents in orthopedics is relatively limited. In 2020, Anthony et al. published the results of a randomized trial demonstrating that using a chatbot to deliver a pain intervention therapy could help limit opioid use post orthopaedic fracture surgery.[7] Bian et al. reported on the use of a AI-assisted conversational agent that mimicked the human voice to follow up patients after orthopedic admission and deliver health education messaging. This study demonstrated considerable cost and time savings compared to their traditional method of follow-up phone calls from health care workers.[8]

One of the key outcome measures in this study was accuracy of responses. Although the minimal acceptable accuracy is unknown, a study by Lin et al. defined accuracy as a >75% accurate response by the chatbot to nonsurvey messaging from patients.[14] In our study, Flix was able to manage nearly 80% of patient questions in an appropriate manner. Felix was only able to correctly identify the topic of the question 58% of the time, and adequately and independently answer the patient question 31% of the time. In many cases of patient interaction Felix was able to self-recognize confusion and appropriately refer to the health care team. These findings are in line with a systematic review of studies using conversational agents performed in 2020, which identified that the most frequently raised issue with chatbots was that the chatbots had difficulty understanding patients because of limited vocabulary.[5]

Other systematic reviews of studies looking at the use of chatbots in healthcare have identified a lack of patient safety outcomes.[6,11] As a result, in our study there was a focus on identifying instances where the chatbot did not recognize or respond appropriately to a serious health concern.[11] We identified 10 questions that were thought to potentially represent patient complications—in three situations of potential DVT or surgical site infections, Felix did not adequately address or recognize the health concern. Although none of these cases eventually represented a true complication, it demonstrates the need to monitor chatbots in health care, especially early in utilization, as well as the continued need to develop and fine-tune the conversational agents.

There is a lot of front-end effort and cost in creating a chatbot that is accurate and augments the patient experience—the cost of developing the Felix chatbot for hip arthroscopy was approximately $USD11,500, which did not include clinician time. Initially, high levels of clinician involvement are required for content development, with a focus on the creation of common questions and appropriate answers. Crowdsourcing from marketplaces, such as Amazon's Mechanical Turk, is then frequently used to generate additional phrasing on a specified topic such as hip arthroscopy, which can increase accuracy.

Limitations

This study has some limitations. First, the timeline of data collection is limited to the first 6 postoperative weeks. However, the first 30 days and, in particular, the first two postoperative weeks are cited as the most vulnerable time post-discharge for complications. Another limitation is the potential generalizability of our study findings to other common orthopedic surgeries, which may have different complications and complications rates, resulting in a different safety profile. The number of patients included in this study is small—the sample size was based on an estimate of 10% incidence of chatbots. Within our study, the frequency of problems with the chatbot was higher than 10% for all evaluated outcomes, making our study well powered to perform analyses. A control group was not

used, so we are not able to determine whether there was a reduction in health care utilization as a result of the use of the chatbot. Finally, we do not have information regarding demographics, such as patient body mass index or the incidence and type of postoperative complications.

Conclusions

The results of this study demonstrate that the use of a chatbot or conversational agent can enhance the postoperative experience for hip arthroscopy patients, as demonstrated by high levels of patient satisfaction.

References

1. Scott EJ, Anthony CA, O'Connor MJ, Lynch TS, Westermann RW. Automated text-messaging after hip arthroscopy: A randomized-controlled trial of "Post-Op Buddy". *Arthroscopy* 2022;38:1488-1495.e1485.
2. Higgins J, Chang J, Hoit G, Chahal J, Dwyer T, Theodoropoulos J. Conventional follow-up versus mobile application home monitoring for postoperative anterior cruciate ligament reconstruction patients: A randomized controlled trial. *Arthroscopy* 2020;36:1906-1916.
3. Scheper H, Derogee R, Mahdad R, et al. A mobile app for postoperative wound care after arthroplasty: Ease of use and perceived usefulness. *Int J Med Infor* 2019;129:75-80.
4. Ponder M, Ansah-Yeboah AA, Charalambous LT, et al. A smartphone app with a digital care pathway for patients undergoing spine surgery: Development and feasibility study. *JMIR Perioper Med* 2020;3:e21138.
5. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: Systematic review. *J Med Internet Res* 2020;22:e20346.
6. Geoghegan L, Scarborough A, Wormald JCR, et al. Automated conversational agents for post-intervention follow-up: A systematic review. *BJS Open* 2021:5.
7. Anthony CA, Rojas EO, Keffala V, et al. Acceptance and commitment therapy delivered via a mobile phone messaging robot to decrease postoperative opioid use in patients with orthopedic trauma: Randomized controlled trial. *J Med Internet Res* 2020;22:e17750.
8. Bian Y, Xiang Y, Tong B, Feng B, Weng X. Artificial intelligence-assisted system in postoperative follow-up of orthopedic patients: Exploratory quantitative and qualitative study. *J Med Internet Res* 2020;22:e16896.
9. Viechtbauer W, Smits L, Kotz D, et al. A simple formula for the calculation of sample size in pilot studies. *J Clin Epidemiol* 2015;68:1375-1379.
10. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: Scoping review and conceptual analysis. *J Med Internet Res* 2020;22:e17158.
11. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: A systematic review. *J Am Med Inform Assoc* 2018;25:1248-1258.
12. Chaix B, Bibault JE, Pienkowski A, et al. When chatbots meet patients: One-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR Cancer* 2019;5:e12856.
13. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: Randomized controlled feasibility trial. *JMIR Mhealth Uhealth* 2019;7:e15018.
14. Lin J, Joseph T, Parga-Belinkie JJ, et al. Development of a practical training method for a healthcare artificial intelligence (AI) chatbot. *BMJ Innovations* 2021;7:441-444.