Comparison of Human-Delivered Conversation versus AI Chatbot Conversation in Increasing Heart Attack Knowledge in Women in the United States

Diane Dagyong Kim, Jingwen Zhang, Kenji Sagae, Holli A. DeVon, Thomas J. Hoffmann, Lauren Rountree, Yoshimi Fukuoka

Submitted to: Journal of Medical Internet Research
on: February 26, 2025

Table of Contents

Comparison of Human-Delivered Conversation versus AI Chatbot Conversation in Increasing Heart Attack Knowledge in Women in the United States

Diane Dagyong Kim[1] MA; Jingwen Zhang[2] PhD; Kenji Sagae[3] PhD; Holli A. DeVon[4] PhD, RN, FAAN, FAHA; Thomas J. Hoffmann[5] PhD; Lauren Rountree[6] RN, PhD; Yoshimi Fukuoka[7] PhD, RN, FAAN

[1] University of California, Davis Davis US
[2]Department of Communication, Department of Public Health Sciences University of California, Davis Davis US
[3]Department of Linguistics University of California, Davis Davis US
[4] University of California, Los Angeles Los Angeles US
[5]Department of Epidemiology & Biostatistics University of California, San Francisco San Francisco US
[6] Massachusetts General Hospital Boston US
[7]Department of Physiological Nursing University of California, San Francisco San Francisco US

Corresponding Author:
Diane Dagyong Kim MA

University of California, Davis
177 Kerr Hall
Davis
US

Abstract

Background: Artificial intelligence (AI) chatbots, driven by advances in natural language processing (NLP), can analyze and generate human language through computational linguistics and machine learning. Despite the rapid development of large language models, little investigation has been conducted to assess whether AI chatbot-delivered educational conversation can achieve a similar level of efficacy as human-delivered conversation.

Objective: To evaluate and compare the potential efficacy of human-delivered conversation versus AI chatbot conversation in increasing women's knowledge and awareness of symptoms and response to heart attack in the United States.

Methods: This is a secondary analysis of two data sets collected from the AI Chatbot Development Project. Women aged 25 years or older were recruited through flyers and social media. The first dataset contained conversational data where a human researcher engaged in educational conversations with women (Human dataset), whereas the second dataset contained conversational data where an AI chatbot named HeartBot engaged in the same educational conversations with women (HeartBot dataset). Knowledge and awareness of symptoms and response to heart attack were measured at the pre-and post-interaction with either the human or HeartBot. Perceived message effectiveness and conversational quality were measured at the post-survey. Ordinal logistic regression analyses were conducted to explore factors predicting women's knowledge, adjusting for age, race/ethnicity, interaction group type, education, word count, message effectiveness, and message humanness.

Results: A total of 171 women (mean age=41 years, SD=12.08) in the Human dataset and 104 women (mean age=46 years, SD=11.86) in the HeartBot dataset completed the baseline survey. Both human-delivered conversations and HeartBot conversation significantly improved participants' ability to recognize heart attack symptoms (AOR 15.19, 95% CI 8.46-27.25, p=0.00000000000000000075; AOR 7.18, 95% CI 3.59-14.36, p=0.000000025), differentiate between symptoms (AOR 9.44, 95% CI 5.60-15.91, p=0.000000000000000034; AOR 5.44, 95% CI 2.76-10.74, p=0.0000011), call emergency services (AOR 6.87, 95% CI 4.09-11.05, p= 0.000000000000035; AOR 5.74, 95% CI 2.84-11.60, p= 0.0000011), and seek emergency care within 60 minutes of symptom onset (AOR 8.68, 95% CI 4.98-15.15, p=0.000000000000027; AOR 2.86, 95% CI 1.55-5.28, p=0.00078) respectively, even after adjusting for covariates. Comparing the two via interaction tests showed a statistically significant improvement of human-delivered conversation vs. HeartBot conversation for all but the calling an ambulance question (p=0.089).

Conclusions: Our findings indicate that HeartBot holds promise in increasing heart attack knowledge and awareness among

women in a cost-effective manner. Future research should employ rigorous experimental designs, such as randomized controlled trials, and evaluate their effectiveness in improving heart health knowledge and subsequent behavior changes.

(JMIR Preprints 26/02/2025:73184)
DOI: https://doi.org/10.2196/preprints.73184

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 v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

JMIR Title Page

Title (147/280 characters)

Comparison of Human-Delivered Conversation versus AI Chatbot Conversation in Increasing Heart Attack Knowledge in Women in the United States

Article type: Original Paper

Authors: Diane Dagyong Kim, Jingwen Zhang, Kenji Sagae, Holli A. DeVon, Thomas J. Hoffmann, Lauren Rountree, Yoshimi Fukuoka

Diane Dagyong Kim, MA
PhD student
Affiliation: Department of Communication, University of California, Davis
Address: Kerr Hall 177, Davis, CA 95616, United States
dagkim@ucdavis.edu
Tel: 650-731-5617
ORCID: 0009-0003-7174-9684

Jingwen Zhang, PhD
Associate Professor
Affiliation: Department of Communication, Department of Public Health Sciences, University of California, Davis
Address: Kerr Hall 362, Davis, CA 95616, United States
jwzzhang@ucdavis.edu
Tel: 530-754-1472
ORCID: 0000-0003-1733-6857

Kenji Sagae, PhD
Professor
Affiliation: Department of Linguistics, University of California, Davis
Address: Kerr Hall 268, Davis, CA 95616, United States
sagae@ucdavis.edu
Tel: 530-754-0998
ORCID: 0000-0003-3371-0618

Holli A. DeVon PhD, RN, FAAN, FAHA
Professor
Affiliation: University of California, Los Angeles
Address: 700 Tiverton Ave, Los Angeles, CA 90095, United States
hdevon@sonnet.ucla.edu
Tel: 310-910-7283
ORCID: 0000 0002 4526 9631

Thomas J. Hoffmann, PhD
Professor
Affiliation: Department of Epidemiology & Biostatistics, University of California, San Francisco
Address: 513 Parnassus Ave, San Francisco, CA 94117, United States
thomas.hoffmann@ucsf.edu
Tel: 415-476-2475
ORCID: 0000-0001-6893-4449

Lauren Rountree, RN, PhD
Staff Nurse
Affiliation: Massachusetts General Hospital
Address: 55 Fruit Street Boston, MA 02114, United States

rountree.lauren@gmail.com
Tel: 978-882-6191
ORCID: 0000-0002-7852-7932

Yoshimi Fukuoka, PhD, RN, FAAN
Professor
Affiliation: Department of Physiological Nursing, University of California, San Francisco
Address: 521 Parnassus Ave, BOX 0638, San Francisco, CA 94143, United States
yoshimi.fukuoka@ucsf.edu
Tel: 415-476-8419
ORCID: 0000-0002-2245-9264

Corresponding author:

Diane Dagyong Kim, MA
PhD student
Affiliation: Department of Communication, University of California, Davis
Address: Kerr Hall 177, Davis, CA 95616, United States
dagkim@ucdavis.edu
Tel: 650-731-5617
ORCID: 0009-0003-7174-9684

Abstract (405/450 words)

Background: Artificial intelligence (AI) chatbots, driven by advances in natural language processing (NLP), can analyze and generate human language through computational linguistics and machine learning. Despite the rapid development of large language models, little investigation has been conducted to assess whether AI chatbot-delivered educational

conversation can achieve a similar level of efficacy as human-delivered conversation.

Objective: To evaluate and compare the potential efficacy of human-delivered conversation versus AI chatbot conversation in increasing women's knowledge and awareness of symptoms and response to heart attack in the United States.

Methods: This is a secondary analysis of two data sets collected from the AI Chatbot Development Project. Women aged 25 years or older were recruited through flyers and social media. The first dataset contained conversational data where a human researcher engaged in educational conversations with women (Human dataset), whereas the second dataset contained conversational data where an AI chatbot named HeartBot engaged in the same educational conversations with women (HeartBot dataset). Knowledge and awareness of symptoms and response to heart attack were measured at the pre-and post-interaction with either the human or HeartBot. Perceived message effectiveness and conversational quality were measured at the post-survey. Ordinal logistic regression analyses were conducted to explore factors predicting women's knowledge, adjusting for age, race/ethnicity, interaction group type, education, word count, message effectiveness, and message humanness.

Results: A total of 171 women (mean age=41 years, SD=12.08) in the Human dataset and 104 women (mean age=46 years, SD=11.86) in the HeartBot dataset completed the baseline survey. Both human-delivered conversations and HeartBot conversation significantly improved participants' ability to recognize heart attack symptoms (AOR 15.19, 95% CI 8.46-27.25, p=0.00000000000000000075; AOR 7.18, 95% CI 3.59-14.36, p=0.000000025), differentiate between symptoms (AOR 9.44, 95% CI 5.60-15.91, p=0.000000000000000034; AOR 5.44, 95% CI 2.76-10.74, p=0.0000011), call emergency services (AOR 6.87, 95% CI 4.09-11.05, p= 0.000000000000035; AOR 5.74, 95% CI 2.84-11.60, p= 0.0000011), and seek emergency care within 60 minutes of symptom onset (AOR 8.68, 95% CI 4.98-15.15, p=0.000000000000027; AOR 2.86, 95% CI 1.55-5.28, p=0.00078) respectively, even after adjusting for covariates. Comparing the two via interaction tests showed a statistically significant improvement of human-delivered conversation vs. HeartBot conversation for all but the calling an ambulance question (p=0.089).

Conclusions: Our findings indicate that HeartBot holds promise in increasing heart attack knowledge and awareness among women in a cost-effective manner. Future research should employ rigorous experimental designs, such as randomized controlled trials, and evaluate their effectiveness in improving heart health knowledge and subsequent behavior changes.

Keywords: artificial intelligence, natural language processing, large language model, chatbot, health education, heart disease, heart attack, knowledge, women

Introduction

AI chatbots built on natural language processing (NLP) techniques represent a prominent application form of artificial intelligence (AI) in healthcare. Chatbots can analyze and generate human language through computational linguistics and machine learning algorithms. Leveraging these capabilities, AI chatbots can process user inputs and assist patients, healthcare providers, and administrators in various ways [1-5]. AI chatbots can engage in natural language conversations and interactions with users through written, oral, and visual communications [6], allowing them to understand and respond to queries, provide information, and offer personalized support through goal setting, counseling, and real-time feedback based on users' personal preferences and behavioral performance data [7]. With 24/7 availability, multilingual support, and ability to handle large volumes of inquiries simultaneously, AI chatbots can enhance healthcare accessibility, efficiency, and patient engagement. AI chatbots have been shown to improve users' health knowledge

acquisition and modify health behaviors [7-10]. In addition, AI chatbots have been built to advance preventive care and mental health services, showing potential to alleviate depression and anxiety, reduce the risk of eating disorders, encourage physical activity, promote healthy diets, and increase self-care behaviors [1,3,8-11].

The fast-growing capabilities of AI chatbots raise questions about their ability to compete with human cognitive and emotional intelligence. Numerous studies have investigated the usability and efficacy of AI chatbots in patients with various health conditions. However, only a few studies have directly compared the effectiveness of AI chatbots to that of human agents. One study revealed that an AI chatbot coach was as effective as a human coach in helping clients reach their goals [12]. Further empirical investigation is needed to more comprehensively evaluate the relative strengths and limitations of AI chatbots in comparison to human agents, especially in the context of healthcare.

Our research team initiated an AI Chatbot Development Project aimed at increasing participants' knowledge and awareness of heart disease. As a first step, we collected a conversational dataset, where a human researcher texted each participant with educational contents on heart health (Human dataset) over two days. We subsequently developed and tested a fully automated SMS-based AI chatbot system named HeartBot, available 24/7, designed to achieve similar objectives, and collected a conversational dataset between HeartBot and participants. The details of findings are under review elsewhere. This project presents a valuable opportunity for a comparative secondary analysis, allowing us to examine the outcomes of the two studies.

The aim of this secondary data analyses was to evaluate and compare the potential efficacy of the two heart disease education interventions. The primary outcome is the knowledge and awareness of symptoms and response to heart attack. In addition, we examined how participant's evaluations on user experience and conversational quality differ between the two formats through assessing message effectiveness, message humanness, naturalness, coherence, and conversational metrics. Our findings contribute to understanding of the relative strengths of human-delivered and automated AI chatbot interventions in health communication, offering practical guidance for designing more effective education and behavior change programs.

Methods

Study Design and Sample

This was a secondary analysis on two data sets collected from the AI Chatbot Development Project conducted from September 2022 to January 2024 [13]. The aims of the AI Chatbot Development Project are to conduct a series of studies to develop a fully automated AI chatbot to increase knowledge and awareness of heart disease in women in the Unites States. After convening a multidisciplinary team, we developed a knowledge bank using the clinical

guidelines, published papers, and American Heart Association “Go Red for Women” materials [14] in order to develop the content of the conversation. Then we conducted a Wizard of Oz experiment with the Human dataset cohort, where participants interacted with a system they believe to be autonomous but was operated by human [15], to test the content and aid in the development of a text-based HeartBot with natural language capabilities. A cardiovascular nurse and PhD student served as the human interventionist to interact with the participants through text-messaging (phase 1: Human dataset).

After the first study, we developed a fully automated AI chatbot, the HeartBot, to deliver the intervention through text-messaging (phase 2: HeartBot dataset). Detailed design of the project, including the protocol, participant eligibility criteria, description of HeartBot platform, and the key usability findings are under review elsewhere [13]. The first phase of the AI Chatbot Development Project was approved by the Institutional Review Board (IRB) at the University of California, Los Angeles (IRB # 22-000878), and the second phase of the project was approved by the IRB at the University of California, San Francisco (IRB # 23-39793).

Eligibility criteria for both studies included women aged 25 years or older who lived in the United States, had access to the internet and a cell phone with texting capabilities, who did not have any self-reported cognitive impairment or history of heart disease or stroke, and who were not health care professionals or students. The eligibility criteria were consistent across the two studies. Participants in the Human dataset were recruited via flyers posted at universities and clinics in Northern and Southern California as well as via advertisement on social media (Facebook and Instagram) from September 2022 to January 2023. Participants in the HeartBot dataset were recruited through Facebook from October 2023 to December 2023.

Measures

Primary Outcomes

Knowledge and Awareness of Symptoms and Response to Heart Attack

To assess the efficacy of a conversational intervention to increase the knowledge and awareness of symptoms and response to heart attack, participants were asked the following four questions on a scale of 1 to 4, in which 1 indicated “not sure” and 4 indicated “sure”: (1) *How sure are you that you could recognize the signs and symptoms of a heart attack in yourself?*, (2) *How sure are you that you could tell the difference between the signs or symptoms of a heart attack and other medical problems?*, (3) *How sure are you that you could call an ambulance or dial 911 if you thought you were having a heart attack?*, (4) *How sure are you that you could get to an emergency room within 60*

minutes after onset of your symptoms of a heart attack?. The same questions were asked before and after interaction with the human researcher and HeartBot. A higher score indicates a better knowledge and awareness of symptoms and response to heart attack.

Other measures

In addition to assessing knowledge and awareness of heart attack as the primary outcome, it is important to examine other measures that can provide insights into user interaction and engagement to the conversation. Evaluating user's experience, conversational quality, and objective conversational metrics such as the number of words used in a conversation allows for comprehensive understanding of how participants perceive and engage with each conversing agent. These insights are essential in identifying areas of improvement and refining HeartBot's design to ensure more effective, natural, and engaging interactions in future iterations.

User Experience

Message Effectiveness

In the AI chatbot behavior change model [16], the effectiveness of chatbot messages is categorized as part of "user experiences", which measures the level of usefulness and convenience in chatbot conversations. Message effectiveness was measured using the *Effectiveness Scale*, a semantic-differential scale originally developed based on previous literature [17, 18]. The scale consists of 5 pairs of opposite adjectives (effective vs. ineffective, helpful vs. unhelpful, beneficial vs. not beneficial, adequate vs. not adequate, supportive vs. not supportive). Participants rated each pair on a 7-point Likert scale, where 1 indicated the negative pole (e.g., "ineffective") and 7 indicated the positive pole (e.g., "effective) in the post-survey. The scores for each item were summed and averaged to create a mean composite score. A higher score indicates greater perceived effectiveness of the messages.

Conversational Quality

Message Humanness

In the AI chatbot behavior change model [16], the humanness of chatbot messages is categorized as part of "conversational quality", which measures the level of humanness and naturalness in chatbot conversations. To evaluate participants' impressions of the messages sent by human researcher and HeartBot, participants rated the humanness of messages using the *Anthropomorphism Scale* [19] in the post-survey. The scale consists of 5 pairs of opposite adjectives (natural vs. fake, humanlike vs. machine-like, conscious vs. unconscious, lifelike vs. artificial, adaptive vs. rigid). Participants rated each pair on a 7-point Likert scale, where 1 indicated the first adjective in the pair (e.g., "natural"), and 7 indicated the second adjective (e.g., "fake"). The scores for each item were summed and averaged to create a mean

composite score. A higher mean score on this measure reflects a more artificial and machine-like impression of the messages.

Conversational Naturalness and Coherence

Conversational quality can be measured from user's subjective evaluation of the conversation's naturalness and coherence as well [16]. To evaluate these, participants were asked to answer the following question in the post-survey: *"Overall, how would you rate the conversations with your texting partner?"*. The response options are: 1) Very unnatural, 2) Unnatural, 3) Neutral, 4) Natural, 5) Very natural. Similarly, participants were also asked to answer the following question in the post-survey: "*Overall, how would you rate the messages you received"*. The response options are: 1) Very incoherent, 2) Incoherent, 3) Neutral, 4) Coherent, 5) Very coherent.

Conversational Metrics

Objective content and linguistic analyses of conversations can be used to evaluate specific dimensions of conversations such as the length of conversations and amount of information exchanged [16]. To measure these dimensions, Linguistic Inquiry and Word Count (LIWC) software [20] was used to process and quantify the total word count of a conversation between the participant and human researcher or HeartBot. The number of words used by each agent (participant, human researcher, HeartBot) was separately measured to process individual contributions within each conversation.

Perception of Chatbot Identity (human versus AI chatbot)

To determine the perception of the identity of the conversing partner, participants were asked the following question at the post-intervention survey: *"Do you think you texted a human or an artificial intelligent chatbot during your conversation?"* There were 2 response options: 1) human, 2) artificial agent.

Covariates: Sociodemographic, past chatbot use, and cardiovascular risks

Sociodemographic measures, such as age, race/ethnicity, education, household income, marital status, and employment status were collected from participants in the baseline survey. In addition, we asked the following question "*Have you used any chatbot in the past 30 days?"* to assess past AI chatbot use experience. The response options were: 1) Yes, 2) No.

The data on self-reported cardiovascular risks, including smoking history, prescribed blood pressure, cholesterol, and diabetes medication intake, and family history of heart disease were collected in the baseline survey. The cardiovascular risk factor variables were selected based on the latest clinical guidelines [21].

Procedure

For the Human dataset, women who were interested in the study were recruited online and underwent screening to confirm eligibility. Eligible participants provided written informed consent prior to enrollment and completed a baseline survey online. Then, participants engaged in two online conversation sessions over the course of two days over a week with a human researcher, with each session covering educational contents related to heart attack symptoms and response. Table 1 presents the content of heart attack topics for women used in both studies. After having a text conversation, participants completed a post survey online to measure knowledge and awareness of symptoms and response to heart attack, message effectiveness, message humanness, conversation naturalness and coherence, and perception of chatbot identity. Participants were provided with a $10 Amazon e-gift card upon completion of all study procedures.

We conducted a follow-up phase developing and evaluating the efficacy of a text-based AI chatbot called HeartBot. HeartBot was programmed using Google Dialogflow and integrated into a text-messaging service called Twilio. Different from human-delivered text conversations, HeartBot engaged one conversation session with the participants. We decided to condense the conversational messaging to only one session to reduce the chance of participant attrition and to make sure participants can receive all educational information within one interaction. In contrast to the first phase of the project (Human dataset), three topics (how angina happens, medicines for heart attack, operational procedures for treating heart attack) were dropped in the second phase (HeartBot dataset), and two quiz questions were included at the end of the conversation to assess participants' retention of key knowledge outcomes. Participants then completed the post survey, and received a $25 Amazon e-gift card. Both studies used the same questionnaires for both the baseline survey and the post survey to measure knowledge and awareness of symptoms and response to heart attack, hosted on a secure online tool called Research Electronic Data Capture (REDCap) [22].

Table 1. Content of heart attack topics for women in the AI Chatbot Development Project.

	Phase 1: Human Dataset	**Phase 2: HeartBot Dataset**
Session 1	1) Greetings 2) What is heart attack 3) Symptoms of heart attacks 4) Leading cause of death for women in the US 5) Gender factors of heart attacks 6) How angina happens 7) Risk factors for heart disease 8) Female specific risk factors for heart disease 9) Racial risk factors of heart disease	1) Greetings 2) Participants' name retrieval 3) Knowledge on heart attacks 4) Symptoms of heart attacks 5) Leading cause of death for women in the US 6) Gender factors of heart attacks 7) First action 8) Importance of calling 911 9) Waiting duration 10) Treatment of heart attacks 11) Action during waiting for 911 12) Risk factors for heart disease

		13) Female specific risk factors for heart disease 14) Racial risk factor for heart disease 15) Multiple choice quiz questions 16) Further questions to ask 17) End of the conversation
Session 2	10) First action 11) Importance of calling 911 12) Waiting duration 13) Tests to diagnose a heart attack 14) Medicines for heart attack 15) Operational procedures for treating heart attack 16) Prevention of heart attacks 17) End of the conversation	Not Applicable

Statistical Analysis

We conducted a descriptive analysis to calculate counts and percentages, or means and standard deviations (SD) for sociodemographic characteristics, past chatbot use, and cardiovascular risks. To compare the two datasets, we performed independent t-tests to assess mean difference for continuous variables and used chi-square tests to examine group distributions. We first conducted Wilcoxon signed-rank tests to evaluate for statistically significant changes in heart attack knowledge and awareness outcome responses (not sure, somewhat not sure, somewhat sure, sure) between the baseline and the post-interaction, within the Human dataset and HeartBot dataset. Then, to adjust for potential confounders, we fit a series of ordinal mixed-effects models using the R v4.1.0 [23] package ordinal v2022.11.16 [24], for each of the four knowledge questions as outcomes. We first fit these models stratified by Human dataset and HeartBot dataset, and adjusting for fixed effects of post (vs. pre; the primary coefficient of interest for these models, indicating whether each of the two interventions was successful), White (vs. non-White), age, interaction group type, education, number of words used by the participants, mean text message effectiveness and humanness of scores, and a random effect for individual. We then fit a model on the entire dataset additionally adjusting for HeartBot (vs. Human), and the interaction between HeartBot and post (i.e., whether HeartBot is more effective than human; the primary coefficient of interest for this model). As an attempted sensitivity analysis, we tried to fit a mixed effects multinomial logistic regression model in Stata v16.1 [25] via the generalized structural equations command, but the models would not converge (likely owing to the small sample size and increased number of parameters to estimate compared to an ordinal logistic regression model). A 2-sided test was used with significance set at $p < .05$.

Results

Sample Characteristics

A total of 171 women in the Human dataset and 104 women in the HeartBot dataset completed the online baseline survey. Table 2 presents the baseline sample characteristics for the two datasets. The mean age (SD) of participants was 41.06 (12.08) years and 46.29 (11.86) years, respectively. In the Human dataset, participants were primarily Black/African American (n=70, 40.9%), college graduates (n=103, 60.3%) and earning moderate to high income (n=68, 39.8%). Participants in the HeartBot dataset were primarily White (n=41, 39.4%), college graduates (n=75, 72.1%) and earning moderate to high income (n=44, 42.3%). A majority of participants in the Human dataset reported having experience in using chatbot (n=96, 56.1%) while the minority of those in the HeartBot dataset did so (n=43, 41.3%).

Table 2. Baseline sample characteristics: sociodemographic, previous chabot use, cardiovascular risks

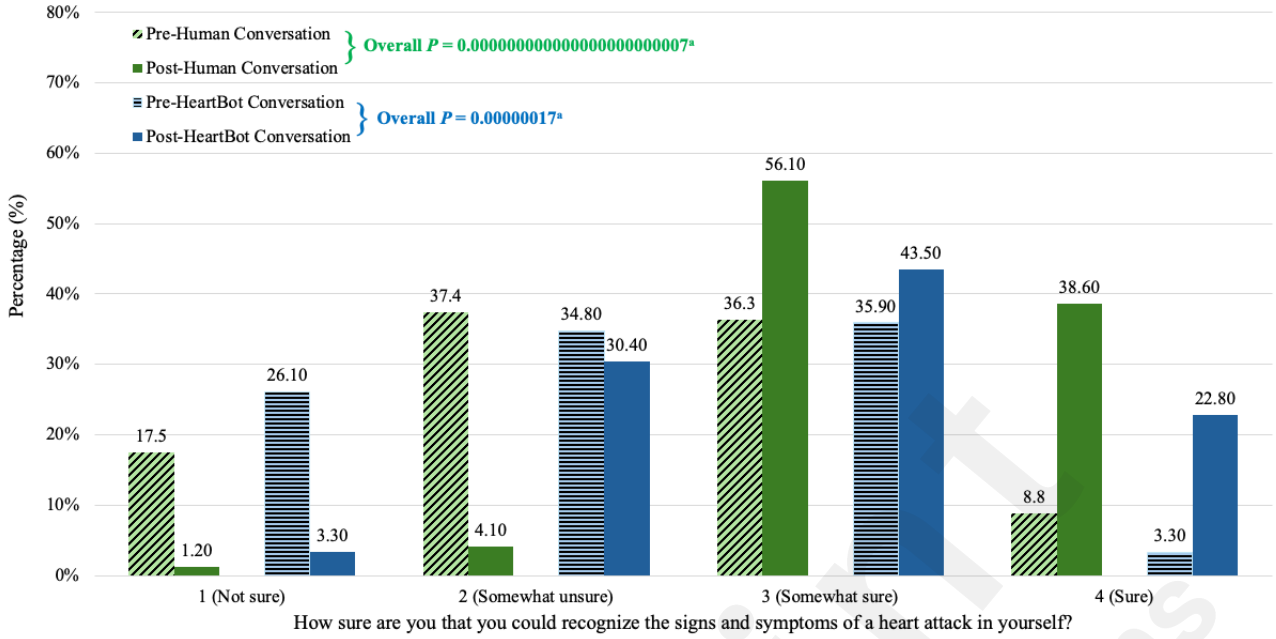
Characteristic	**Human Dataset (N=171)** Mean (SD)/[range] or n (%)	**HeartBot Dataset (N=104)** Mean (SD)/[range] or n (%)	*P* value
Sociodemographic			
Age (years)	41.06 (12.08) / [25.0-76.0]	46.29 (11.86) / [26.0-70.0]	<.001
Race/Ethnicity			0.018
Black/African American (non-Hispanic)	70 (40.9)	22 (21.2)	
Hispanic/Latino	29 (17.0)	26 (25.0)	
Asian	10 (5.8)	6 (5.8)	
White (non-Hispanic)	50 (29.2)	41 (39.4)	
American Indian/ Native Hawaiian/ More than 1 race/ethnicity	12 (7.0)	9 (8.7)	
Education			0.046
Completed some college course work, but did not finish or less	68 (39.8)	29 (27.9)	
Completed college/graduate school	103 (60.3)	75 (72.1)	
Household income			
Less than $40,000/Don’t know	59 (34.5)	27 (26.0)	0.295
$40,001 - $75,000	44 (25.7)	33 (31.7)	
Greater than $75,000	68 (39.8)	44 (42.3)	
Marital status			0.654
Never married	46 (26.9)	23 (22.1)	
Currently married/cohabitating	108 (63.2)	69 (66.3)	
Divorced/widowed	17 (9.9)	12 (11.5)	
Employment status			0.074
Full-time/Part-time	108 (63.2)	62 (59.6)	
Unemployed/Homemaker/ Student	42 (24.5)	19 (18.3)	
Retired/Disabled/Other	21 (12.3)	23 (22.1)	
Chatbot use history			

Chatbot use (e.g., Amazon's Alexa, Google Assistant, Siri, Facebook Messenger bot etc.) in the past 30 days			0.683
Yes	96 (56.1)	43 (41.3)	
No	75 (43.9)	61 (58.7)	
Cardiovascular risks			
Smoked at least one cigarette in the last 30 days			0.063
Yes	14 (8.2)	16 (15.4)	
No	157 (91.8)	88 (84.6)	
Blood pressure medication			0.077
Yes	71 (41.5)	30 (28.8)	
No/Don’t know	100 (58.5)	74 (71.2)	
Cholesterol medication			0.711
Yes	62 (36.3)	33 (31.7)	
No/Don’t know	109 (63.8)	71 (68.3)	
Diabetes medication			0.841
Yes	23 (13.5)	13 (12.5)	
No/Don’t know	148 (86.6)	91 (87.5)	
Family history of heart disease/stroke			0.237
Yes	38 (22.2)	17 (16.3)	
No/Don’t know	133 (77.8)	87 (83.7)	

Changes in Knowledge and Awareness of Heart Disease

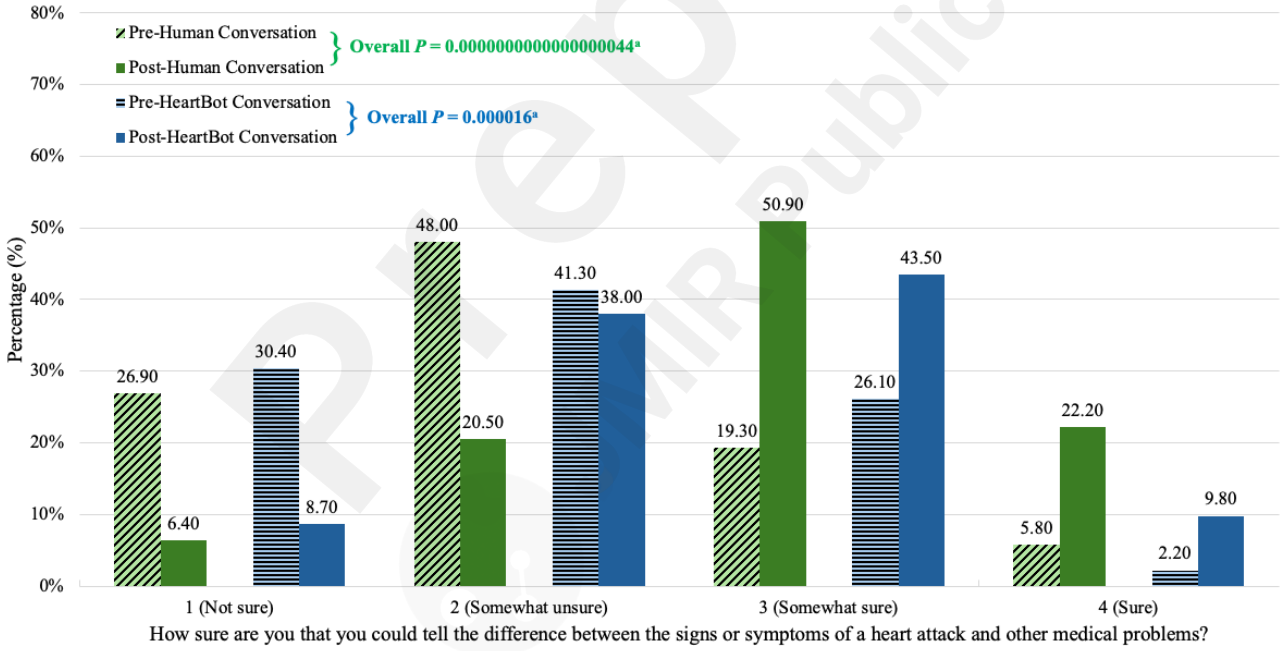
Figure 1-4 presents the results of Wilcoxon signed-rank tests for the four outcomes in changes in knowledge and awareness of heart disease. Table Supplement 1 also shows the detailed result of the Figure 1-4. Overall, Wilcoxon signed-rank tests revealed significant increase in knowledge and awareness of heart disease across all four outcome measures following interactions with both human researcher and HeartBot ($p<0.05$).

Figure 1. Change in response between pre and post Human conversation (n=171), and pre and post HeartBot conversation (n=92) for knowledge on recognizing signs and symptoms of a heart attack.



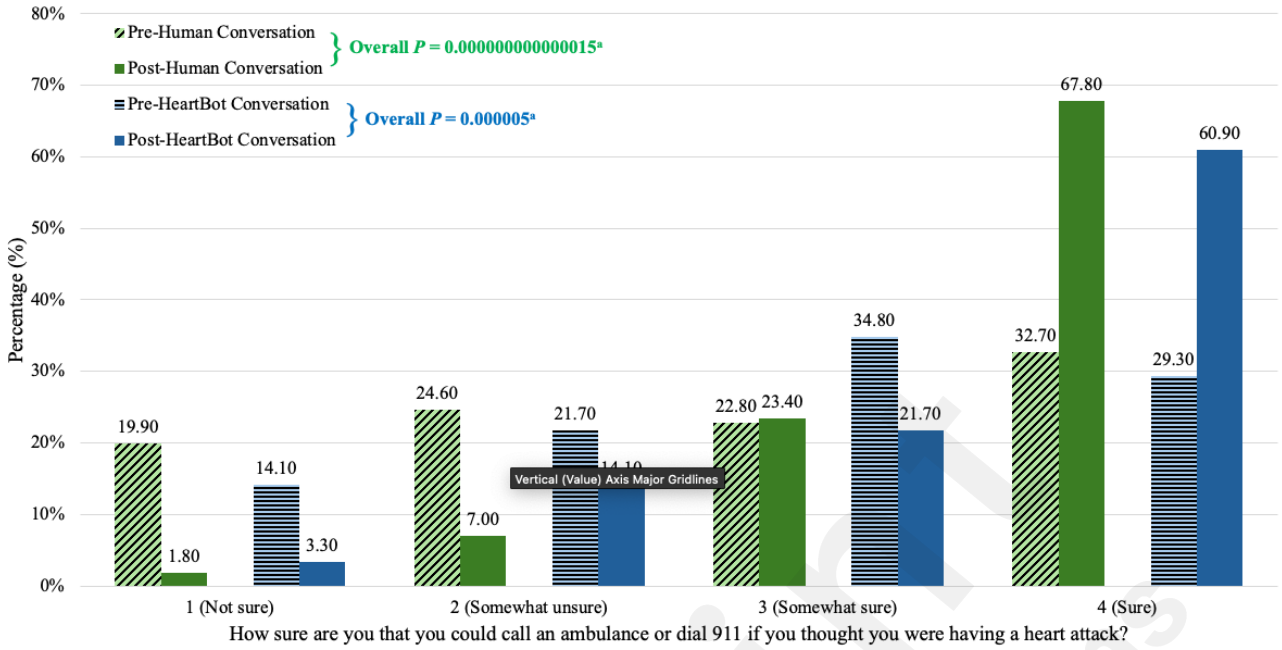
aWilcoxon matched pairs test was conducted.

Figure 2. Change in response between pre and post Human conversation (n=171), and pre and post HeartBot conversation (n=92) for knowledge on telling the difference between the signs or symptoms of a heart attack and other medical problems.



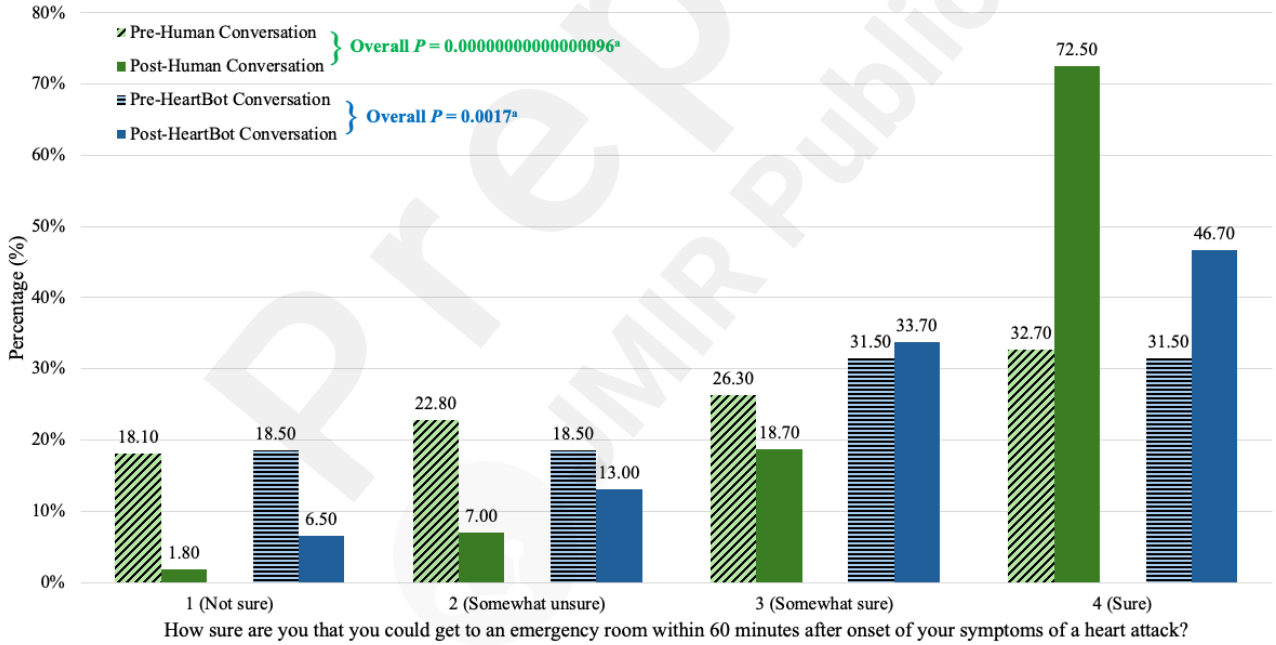
aWilcoxon matched pairs test was conducted.

Figure 3. Change in response between pre and post Human conversation (n=171), and pre and post HeartBot conversation (n=92) for knowledge on calling an ambulance or dialing 911 when experiencing heart attack.



[a]Wilcoxon matched pairs test was conducted.

Figure 4. Change in response between pre and post Human conversation (n=171), and pre and post HeartBot conversation (n=92) for knowledge on getting to an emergency room within 60 minutes after onset of symptoms of a heart attack.



[a]Wilcoxon matched pairs test was conducted.

Table 3 shows the adjusted odds ratios from a series of ordinal logistic regression analyses for predicting each knowledge question for the Human dataset. In the Human dataset, after controlling for age, ethnicity, education, message effectiveness, message humanness, and chatbot use history, the human-delivered conversations improved participants' knowledge and awareness in recognizing the signs and symptoms of a heart attack response (AOR 15.19, 95% CI 8.46-27.25, p=0.00000000000000000075), telling the difference between the signs or symptoms of a heart attack response

(AOR 9.44, 95% CI 5.60-15.91, p=0.0000000000000000034), calling an ambulance or dialing 911 during heart attack response (AOR 6.87, 95% CI 4.09-11.05, p=0.00000000000035), and getting to an emergency room within 60 minutes after onset of symptoms response (AOR 8.68, 95% CI 4.98-15.15, p=0.000000000000027). In the HeartBot dataset, these effects were generally reduced but still substantially improved (see Table 3, full model in Table Supplement 2), e.g., in the recognizing the signs and symptoms questions (AOR 7.18, 95% CI 3.59-14.36, p=0.000000025). A formal interaction test showed a statistically significant improvement of Human vs. HeartBot dataset for all but the third question (calling an ambulance; P=0.089; Table 3, Table Supplement 2). We could not adjust for word count, as all human-delivered conversations in the Human dataset were longer than any of the HeartBot conversations in the HeartBot dataset, and so the model would not fit.

Table 3. Ordinal logistic regression models comparing post- vs. pre-intervention (Human or HeartBot) on the four knowledge questions.[a]

Cohort	Term	Q1: Recognizing signs and symptoms of a heart attack		Q2: Telling the difference between the signs or symptoms of a heart attack and other medical problems		Q3: Calling an ambulance or dialing 911 when experiencing heart attack		Q4: Getting to an emergency room within 60 minutes after onset of symptoms of a heart attack	
		AOR (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value
Human-Delivered Text Conversation	Post (vs. pre)	15.19 (8.46, 27.25)	0.00000000000000075	9.44 (5.60, 15.91)	0.0000000000000034	6.87 (4.09, 11.55)	0.00000000035	8.68 (4.98, 15.15)	0.000000000000027
HeartBot Conversation	Post (vs. pre)	7.18 (3.59, 14.36)	0.000000025	5.44 (2.76, 10.74)	0.0000011	5.74 (2.84, 11.60)	0.0000011	2.86 (1.55, 5.28)	0.00078
All	Post x Heartbot	0.38 (0.19, 0.78)	0.0084	0.40 (0.20, 0.80)	0.01	0.53 (0.25, 1.10)	0.089	0.26 (0.12, 0.55)	0.00044

[a]Models are additionally adjusted for White (vs. non-White), age, group type, education, user word count, mean text message effectiveness, and humanness of scores (full models in Table S2); Q1, How sure are you that you could recognize the signs and symptoms of a heart attack in yourself? (Select a number from 1: not sure to 4: sure); Q2: How sure are you that you could tell the difference between the signs or symptoms of a heart attack and other medical problems? (Select a number from 1: not sure to 4: sure); Q3: How sure are you that you could call an ambulance or dial 911 if you thought you were having a heart attack? (Select a number from 1: not sure to 4: sure); Q4, How sure are you that you could get to an emergency room within 60 minutes after onset of your symptoms? (Select a number from 1: not sure to 4: sure); *, P<0.05; **, P<0.01; ***, P<0.001; Abbreviation: AOR, adjusted odds ratio; 95% CI, 95% confidence interval.

Human-Delivered Conversation versus HeartBot Conversation

Table 4 presents the comparison of the evaluation of conversation quality between the two studies. In the

Human dataset, participants interacted with the human researcher and completed conversation sessions over the course of two days. The mean (SD) and median number of words used by the participants and their conversing agent overall were 2322 (875.65) and 2097 words in the Human dataset, and 888.04 (76.04) and 852 words in the HeartBot dataset. Participants in the Human dataset ranked all conversational qualities, which include message effectiveness, message humanness, conversation naturalness and coherence, significantly higher than those in the HeartBot dataset. The majority of participants in both groups correctly identified in they were conversing with a human or Heartbot.

Table 4. Comparing the evaluation of conversation quality between Human Dataset and HeartBot Dataset.

	Human Dataset (n=171) Mean (SD)/[range] or n (%)	**HeartBot Dataset (n=92)** Mean (SD)/[range] or n (%)	***P* value**
User Experience			
Score of Message Effectiveness scale	6.35 (0.85)	5.66 (1.23)	<.001
Conversation Quality (subjective measure)			
Score of Message Humanness scale	5.86 (1.24)	5.19 (1.19)	<.001
Overall, how would you rate the conversations with your texting partner?			<.001
Very unnatural/Unnatural	9 (5.3)	5 (5.4)	
Neutral	19 (11.1)	33 (35.9)	
Natural/Very natural	143 (83.6)	54 (58.7)	
Overall, how would you rate the messages you received?			<.001
Very incoherent/Incoherent	1 (0.6)	0 (0)	
Neutral	4 (2.3)	23 (25.0)	
Coherent/Very coherent	143 (83.6)	69 (75.0)	
Conversation Quality (Objective measure)			
Number of words used by the participants and human researcher/HeartBot	2322.55 (875.65)/[1314.0-8073.0] Median: (2097.0)	888.04 (76.4)/[778-1274]/ Median: (852)	<.001
Number of words used by the participants	298.94 (227.90)/[83.0-1986.0] Median: (231.0)	80.57 (60.19)/[34-377] Median: (63)	<.001
Do you think you texted a human or artificial intelligent chatbot during your conversation?			<.001
Human	127 (74.3)	31 (33.7)	
Artificial Intelligent Chatbot	44 (25.7)	61 (66.3)	

Discussion

Principal Results

We compared the potential efficacy of human-delivered conversation versus AI chatbot conversation in increasing women's knowledge and awareness of symptoms and the appropriate response to heart attack in the United States, while controlling for potential confounding factors. Since this study was designed to be exploratory and not based on a randomized control trial, true efficacy cannot be established. The results reveal that interacting with both the human and HeartBot were significantly effective in increasing knowledge and awareness of heart attack among participants (i.e. recognizing signs and symptoms of a heart attack, telling the difference between the signs or symptoms of a heart attack

and other medical problems, calling an ambulance or dialing 911 when experiencing heart attack, getting to an emergency room within 60 minutes after onset of symptoms of a heart attack). However, human-delivered conversation had a greater efficacy than HeartBot conversation for all except the calling an ambulance question (p=0.089).

Several potential explanations can be considered due to the inherent differences in the content and duration of the conversation sessions. First, human-delivered conversation involved a more extended engagement process, comprising two separate sessions over a week, allowing participants to engage in a more prolonged and reflective learning process. In contrast, the HeartBot conversation was limited to a single session, which may have constrained the depth of discussion. Second, participants in the Human dataset produced significantly more words during the conversation, with a mean (SD) word count of 298.94 (875.65), compared to 80.57 (60.19) in the HeartBot conversation. The greater verbosity in the Human dataset may have contributed to deeper discussions and enhanced knowledge reinforcement, potentially explaining the observed increase in efficacy. However, we were not able to statistically account for word count, as models adjusting for the covariate would not converge, likely owing to having very different distributions of word counts with little overlap in the two groups (humans a mean [SD] of 2322 [875.65] words, HeartBot a mean [SD] of 888.04 [76.04] words). Finally, human-delivered conversations were facilitated by a human researcher, who is a nurse, allowing for greater flexibility in language use, response adaptation, and addressing participant queries in a more personalized manner. In contrast, HeartBot had the inherent limitation in conversational algorithm, which appears less personalized and less flexible, followed a structured script, limiting its ability to adjust dynamically to participants' specific concerns.

HeartBot, fully automated AI chatbot, did significantly increase participants' knowledge and awareness to symptoms and response to heart attack, and demonstrates significant potential as an innovative digital health intervention. As traditional public health campaigns struggle to engage diverse populations, AI chatbots offer a scalable, 24/7 accessible, and personalized approach to health education for broader populations. Their adaptive algorithms allow for dynamic personalization, tailoring responses to individual user queries and comprehension levels, which may enhance engagement and knowledge retention beyond one-size-fits-all campaigns. Additionally, chatbot interactions require active engagement as participants read, process, and respond to information, reinforcing learning through interaction rather than passive intake [26]. The anonymized nature of chatbot conversations can also reduce psychological barriers, encouraging users to seek information more openly, especially on sensitive health topics [27]. Finally, HeartBot integrates structured quiz components, encouraging reinforcement of learning through immediate self-assessment and cognitive recall. While these advantages highlight AI chatbot's potential, findings from this study suggest room for improvement to further

enhance its efficacy. First, increasing the number of interaction sessions—rather than a single one-time interaction—may allow for more sustained engagement and deeper knowledge retention, aligning more closely with multi-session format of human-delivered conversations. Second, further iterations could leverage machine learning algorithms to continuously refine conversation models and improve HeartBot's flexibility in answering participants' queries, which could make interaction with HeartBot feel more responsive and personalized. Lastly, to fully evaluate HeartBot's long-term efficacy and potential parity with human-delivered conversations, a randomized controlled trial (RCT) would be instrumental.

Interestingly, although we did not disclose the identity of the conversing partner, participants showed misperception regarding the identity of their conversing partner across both conversation phases. In the Human dataset, 25.7% of participants believed they were interacting with an AI chatbot, while 33.7% of participants in the HeartBot dataset perceived their partner as human. This misattribution suggests a certain level of ambiguity in perceived conversational agency. Moreover, the mean scores of message effectiveness and message humanness scales were high in both conversation conditions (above 3.5 on a 1-7 scale), but higher in the Human dataset, with means (SD) of 6.35 (0.85) and 5.86 (1.24), respectively, compared to 5.66 (1.23) and 5.19 (1.19) in the HeartBot dataset. Similarly, participants rated both conversation conditions highly for naturalness and coherence. However, those in the Human dataset perceived the conversation as more coherent (83.6%) and natural (83.6) compared to participants in the HeartBot dataset, with coherence rated at 58.7% and naturalness at 75.0%. This distinction may be attributed to the inherent challenge of replicating human communication subtleties in algorithmic interactions. While HeartBot demonstrated considerable communicative competence, it encountered limitations in fully imitating the nuanced relational aspects of human dialogue. Drawing from the Computers Are Social Actors (CASA) paradigm [28], participants apply social interaction schemas to technological interfaces, yet experience these interactions with less emotional depth and relational intimacy. Key communication studies have consistently highlighted the critical role of relational cues in establishing trust and engagement and promote human-chatbot relationships [29]. For example, research has shown that conversational agents can build positive relationships in health and well-being setting through verbal behaviors like humor [29], social dialogue [30], and empathy [31. Although HeartBot successfully delivered equivalent factual content, it inherently struggled to reproduce the affective dimensions that characterize human-to-human communication. These findings suggest that while AI chatbots provide a promising technological intervention, they must continue to evolve in their ability to simulate the nuanced relational components of effective human health communication.

Strengths and limitations

This study is among the few that directly compare the impact of Human-Delivered Text Conversation and

HeartBot Conversation on knowledge and awareness of heart disease outcomes in community-dwelling women in the United States. By examining both approaches in designing and improving the efficacy of the fully automated AI chatbot, we address a gap in the literature where few studies explore relative impact of human versus automated AI chatbot interventions. Our findings highlight that both approaches significantly increased contribute positively to knowledge and awareness enhancement, suggesting that fully automated AI chatbots have great potential to educate women and improve public knowledge and awareness of heart disease. In addition, the sample in this study represents a wide range of women from diverse backgrounds, including individuals with no prior experience in using chatbots and a large number of women with racial/ethnic minority backgrounds. Representation of a wide range of participants enhances the generalizability of the study's results.

Several limitations of the study need to be acknowledged. First, as this study was not designed as a randomized controlled trial, we are unable to establish a causal relationship between the intervention and knowledge outcomes. Second, this study used a convenience sample, which might introduce selection bias, as participants were self-selected women willing to participate in the technology-based interventions.

Conclusion

Our findings suggest that fully automated AI chatbots (HeartBot) hold significant potential to improve knowledge and awareness of heart attack symptoms and response in women, presenting a scalable option for public health interventions. With the capacity to reach a broad audience at relatively low cost, chatbots offer a promising opportunity for delivering accessible health education to diverse populations. However, to rigorously assess their efficacy, future research should incorporate randomized controlled trials to establish causality and evaluate the AI chatbot's effectiveness in knowledge improvement. Such studies could provide more definitive evidence on the value of AI chatbot-based health interventions and guide the development of optimized digital health tools for public use.

Acknowledgements

The project was supported by the Noyce Foundation and the UCSF School of Nursing Gain Fund. The project sponsors had no role in the study design, collection, analysis, or interpretation of data, writing the report, or deciding to submit the report for publication.

Author's Contributions

YF is the principal investigator and obtained research funding for this project. DK, JZ, KS, HAD, TJH, LR and YF contributed to the conception and design. DK, TJH and YF wrote sections of the manuscript. DK, LR and YF contributed to collecting all data. DK, TJH and YF contributed to statistical analysis. All authors reviewed the manuscript

and approved the submitted version.

Conflicts of Interest

None declared.

Abbreviations

AI: artificial intelligence, AOR: adjusted odds ratio

Table Supplement 1. Change in women's knowledge and awareness of symptoms and response to heart attack between pre and post Human interaction (n=171), and pre and post HeartBot interaction (n=92) on the four knowledge questions.

	Human Data (n=171)			HeartBot Data (n=92)		
	Pre-survey	Post-survey	Wilcoxon *P* value	Pre-survey	Post-survey	Wilcoxon *P* value
Responding to a Potential Heart Attack						
How sure are you that you could recognize the signs and symptoms of a heart attack in yourself? (Please select a number from 1-4)						
1: Not sure	30 (17.5)	2 (1.2)	<.001	24 (26.1)	3 (3.3)	<.001
2	64 (37.4)	7 (4.1)		32 (34.8)	28 (30.4)	
3	62 (36.3)	96 (56.1)		33 (35.9)	40 (43.5)	
4: Sure	15 (8.8)	66 (38.6)		3 (3.3)	21 (22.8)	
Mean (SD)	2.36 (0.87)	3.32 (0.61)		2.16 (0.86)	2.86 (0.81)	<.001
1, 2: not sure	94 (55.0)	9 (5.3)	<.001	56 (60.9)	31 (33.7)	<.001
3, 4: sure	77 (45.0)	162 (94.7)		36 (39.1)	61 (66.3)	
How sure are you that you could tell the difference between the signs or symptoms of a heart attack and other medical problems? (Please select a number from 1-4)						
1: Not sure	46 (26.9)	11 (6.4)	<.001	28 (30.4)	8 (8.7)	<.001
2	82 (48.0)	35 (20.5)		38 (41.3)	35 (38.0)	
3	33 (19.3)	87 (50.9)		24 (26.1)	40 (43.5)	
4: Sure	10 (5.8)	38 (22.2)		2 (2.2)	9 (9.8)	
Mean (SD)	2.04 (0.84)	2.89 (0.82)		2.00 (0.81)	2.54 (0.79)	<.001
1, 2: not sure	128 (74.9)	46 (26.9)	<.001	66 (71.7)	43 (46.7)	<.001
3, 4: sure	43 (25.1)	125 (73.1)		26 (28.3)	49 (53.3)	
How sure are you that you could call an ambulance or dial 911 if you thought you were having a heart attack? (Please select a number from 1-4)						
1: Not sure	34 (19.9)	3 (1.8)	<.001	13 (14.1)	3 (3.3)	0.018
2	42 (24.6)	12 (7.0)		20 (21.7)	13 (14.1)	
3	39 (22.8)	40 (23.4)		32 (34.8)	20 (21.7)	
4: Sure	56 (32.7)	116 (67.8)		27 (29.3)	56 (60.9)	
Mean (SD)	2.68 (1.13)	3.57 (0.70)		2.79 (1.02)	3.40 (0.85)	<.001
1, 2: not sure	76 (44.4)	15 (8.8)	<.001	33 (35.9)	16 (17.4)	<.001
3, 4: sure	95 (55.6)	156 (91.2)		59 (64.1)	76 (82.6)	
How sure are you that you could get to an emergency room within 60 minutes after onset of your symptoms of a heart attack? (Please select a number from 1-4)						
1: Not sure	31 (18.1)	3 (1.8)	<.001	17 (18.5)	6 (6.5)	0.002
2	39 (22.8)	12 (7.0)		17 (18.5)	12 (13.0)	
3	45 (26.3)	32 (18.7)		29 (31.5)	31 (33.7)	
4: Sure	56 (32.7)	124 (72.5)		29 (31.5)	43 (46.7)	
Mean (SD)	2.74 (1.10)	3.62 (0.70)		2.76 (1.09)	3.21 (0.91)	<.001
1, 2: not sure	70 (40.9)	15 (8.8)	<.001	34 (37.0)	18 (19.6)	<.001
3, 4: sure	101 (59.1)	156 (91.2)		58 (63.0)	74 (80.4)	

Table Supplement 2. Full ordinal logistic regression models.

(a) Human Text Conversation

Variable	Q1: Recognizing signs and symptoms of a heart attack		Q2: Telling the difference between the signs or symptoms of a heart attack and other medical problems		Q3: Calling an ambulance or dialing 911 when experiencing heart attack		Q4: Getting to an emergency room within 60 minutes after onset of symptoms of a heart attack	
	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value
P: Post (vs. pre) knowledge	15.19 (8.46, 27.25)	7.5e-20***	9.44 (5.60, 15.91)	3.4e-17***	6.87 (4.09, 11.55)	3.5e-13***	8.68 (4.98, 15.15)	2.7e-14***
Age (years)	0.99 (0.97, 1.01)	0.44	0.99 (0.97, 1.01)	0.32	1.01 (0.99, 1.03)	0.32	1.02 (1.00, 1.05)	0.059
Non-white (vs. white)	0.90 (0.52, 1.58)	0.72	1.30 (0.71, 2.40)	0.4	1.36 (0.77, 2.41)	0.3	0.50 (0.26, 0.98)*	0.045
Education (completed collage/graduate school vs. not)	0.64 (0.37, 1.09)	0.098	0.58 (0.32, 1.03)	0.064	0.63 (0.36, 1.11)	0.11	0.72 (0.39, 1.32)	0.29
Message effectiveness[a]	0.84 (0.57, 1.25)	0.4	0.84 (0.55, 1.30)	0.44	0.95 (0.63, 1.43)	0.81	0.98 (0.63, 1.54)	0.94
Message humanness[a]	1.36 (1.03, 1.79)*	0.029	1.33 (0.99, 1.80)	0.061	1.19 (0.90, 1.58)	0.23	0.98 (0.72, 1.34)	0.9
Chatbot use history	0.99 (0.62, 1.60)	0.98	0.95 (0.57, 1.60)	0.85	0.80 (0.49, 1.33)	0.4	0.96 (0.55, 1.67)	0.89

Footnote. Q1, How sure are you that you could recognize the signs and symptoms of a heart attack in yourself? (Select a number from 1: not sure to 4: sure); Q2: How sure are you that you could tell the difference between the signs or symptoms of a heart attack and other medical problems? (Select a number from 1: not sure to 4: sure); Q3: How sure are you that you could call an ambulance or dial 911 if you thought you were having a heart attack? (Select a number from 1: not sure to 4: sure); Q4, How sure are you that you could get to an emergency room within 60 minutes after onset of your symptoms? (Select a number from 1: not sure to 4: sure); *, P<0.05; **, P<0.01; ***, P<0.001. [a]Message effectiveness and message humanness each consist of 5 items, and each item scores are based on a 7-point Likert scale. The scores of the 5 items for message effectiveness and message impression were summed and averaged to create a mean composite score (Cronbach α=0.94 and 0.91, respectively); Abbreviation: AOR, adjusted odds ratio; 95% CI, 95% confidence interval.

(b) HeartBot Conversation

Variable	Q1: Recognizing signs and symptoms of a heart attack		Q2: Telling the difference between the signs or symptoms of a heart attack and other medical problems		Q3: Calling an ambulance or dialing 911 when experiencing heart attack		Q4: Getting to an emergency room within 60 minutes after onset of symptoms of a heart attack	
	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value
P: Post (vs. pre) knowledge	7.18 (3.59, 14.36)	2.5e-08***	5.44 (2.76, 10.74)	1.1e-06***	5.74 (2.84, 11.60)	1.1e-06***	2.86 (1.55, 5.28)	0.00078***
Age (years)	0.97 (0.93, 1.00)	0.055	0.97 (0.94, 1.01)	0.16	0.97 (0.94, 1.01)	0.16	1.04 (1.00, 1.07)	0.044*
Education (completed	0.35 (0.14, 0.87)	0.025*	0.19 (0.07, 0.51)	0.00099***	0.56 (0.21, 1.48)	0.24	0.37 (0.15, 0.93)	0.035*

collage/graduate school vs. not)								
Message effectiveness[a]	1.60 (1.03, 2.48)	0.037*	1.62 (1.02, 2.59)	0.042*	1.81 (1.09, 3.00)	0.022*	1.38 (0.90, 2.11)	0.14
Message humanness[a]	1.20 (0.77, 1.87)	0.42	1.06 (0.67, 1.68)	0.81	0.91 (0.55, 1.53)	0.73	1.24 (0.79, 1.94)	0.00078
Chatbot use history	1.32 (0.60, 2.87)	0.49	1.71 (0.75, 3.87)	0.2	0.89 (0.38, 2.09)	0.8	0.86 (0.40, 1.86)	0.044

Footnote. Q1, How sure are you that you could recognize the signs and symptoms of a heart attack in yourself? (Select a number from 1: not sure to 4: sure); Q2: How sure are you that you could tell the difference between the signs or symptoms of a heart attack and other medical problems? (Select a number from 1: not sure to 4: sure); Q3: How sure are you that you could call an ambulance or dial 911 if you thought you were having a heart attack? (Select a number from 1: not sure to 4: sure); Q4, How sure are you that you could get to an emergency room within 60 minutes after onset of your symptoms? (Select a number from 1: not sure to 4: sure); *, P<0.05; **, P<0.01; ***, P<0.001. [a]Message effectiveness and message humanness each consist of 5 items, and each item scores are based on a 7-point Likert scale. The scores of the 5 items for message effectiveness and message impression were summed and averaged to create a mean composite score (Cronbach α=0.94 and 0.91, respectively); Abbreviation: AOR, adjusted odds ratio; 95% CI, 95% confidence interval.

(c) All data

Variable	**Q1: Recognizing signs and symptoms of a heart attack**		**Q2: Telling the difference between the signs or symptoms of a heart attack and other medical problems**		**Q3: Calling an ambulance or dialing 911 when experiencing heart attack**		**Q4: Getting to an emergency room within 60 minutes after onset of symptoms of a heart attack**	
	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value	AOR (95% CI)	*P* value
P: Post (vs. pre) knowledge	16.29 (9.40, 28.20)	2.3e-23***	11.08 (6.68, 18.37)***	1.1e-20***	8.25 (4.91, 13.86)***	1.5e-15***	9.79 (5.72, 16.75)	8.2e-17***
H: Heartbot vs. human	0.97 (0.53, 1.79)	0.92	1.25 (0.66, 2.37)	0.5	1.38 (0.75, 2.53)	0.3	1.45 (0.77, 2.71)	0.25
P x H interaction	0.38 (0.19, 0.78)	0.0084**	0.40 (0.20, 0.80)	0.01*	0.53 (0.25, 1.10)	0.089	0.26 (0.12, 0.55)	0.00044***
Age (years)	0.98 (0.97, 1.00)	0.084	0.99 (0.97, 1.01)	0.17	1.00 (0.98, 1.02)	0.91	1.03 (1.01, 1.05)	0.0045**
Non-white (vs. white)	0.85 (0.47, 1.55)	0.6	1.17 (0.62, 2.20)	0.62	1.47 (0.80, 2.71)	0.22	0.41 (0.21, 0.79)	0.0082**
Education (completed collage/graduate school vs. not)	0.55 (0.34, 0.89)	0.015*	0.42 (0.26, 0.70)	0.00075***	0.64 (0.39, 1.05)	0.08	0.58 (0.35, 0.96)	0.034*
Message effectiveness[a]	1.21 (0.91, 1.60)	0.19	1.17 (0.86, 1.58)	0.31	1.30 (0.96, 1.75)	0.085	1.24 (0.92, 1.68)	0.16
Message humanness[a]	1.23 (0.97, 1.55)	0.087	1.17 (0.91, 1.50)	0.22	1.06 (0.83, 1.36)	0.64	0.99 (0.77, 1.28)	0.97
Chatbot use history	1.02 (0.67, 1.54)	0.94	1.09 (0.70, 1.69)	0.71	0.77 (0.50, 1.19)	0.24	0.89 (0.57, 1.39)	0.61

Footnote. Q1, How sure are you that you could recognize the signs and symptoms of a heart attack in yourself? (Select a number from 1: not sure to 4: sure); Q2: How sure are you that you could tell the difference between the signs or symptoms of a heart attack and other medical problems? (Select a number from 1: not sure to 4: sure); Q3: How sure are you that you could call an ambulance or dial 911 if you thought you were having a heart attack? (Select a number from 1: not sure to 4: sure); Q4, How sure are you that you could get to an emergency room within 60 minutes after onset of your symptoms? (Select a number from 1: not sure to 4: sure); *, P<0.05; **, P<0.01; ***, P<0.001. [a]Message effectiveness and message humanness each consist of 5 items, and each item scores are based on a 7-point Likert scale. The scores of the 5 items for message effectiveness and message impression were summed and averaged to create a mean composite score (Cronbach α=0.94 and 0.91, respectively); Abbreviation: AOR, adjusted odds ratio; 95% CI, 95% confidence interval.

References

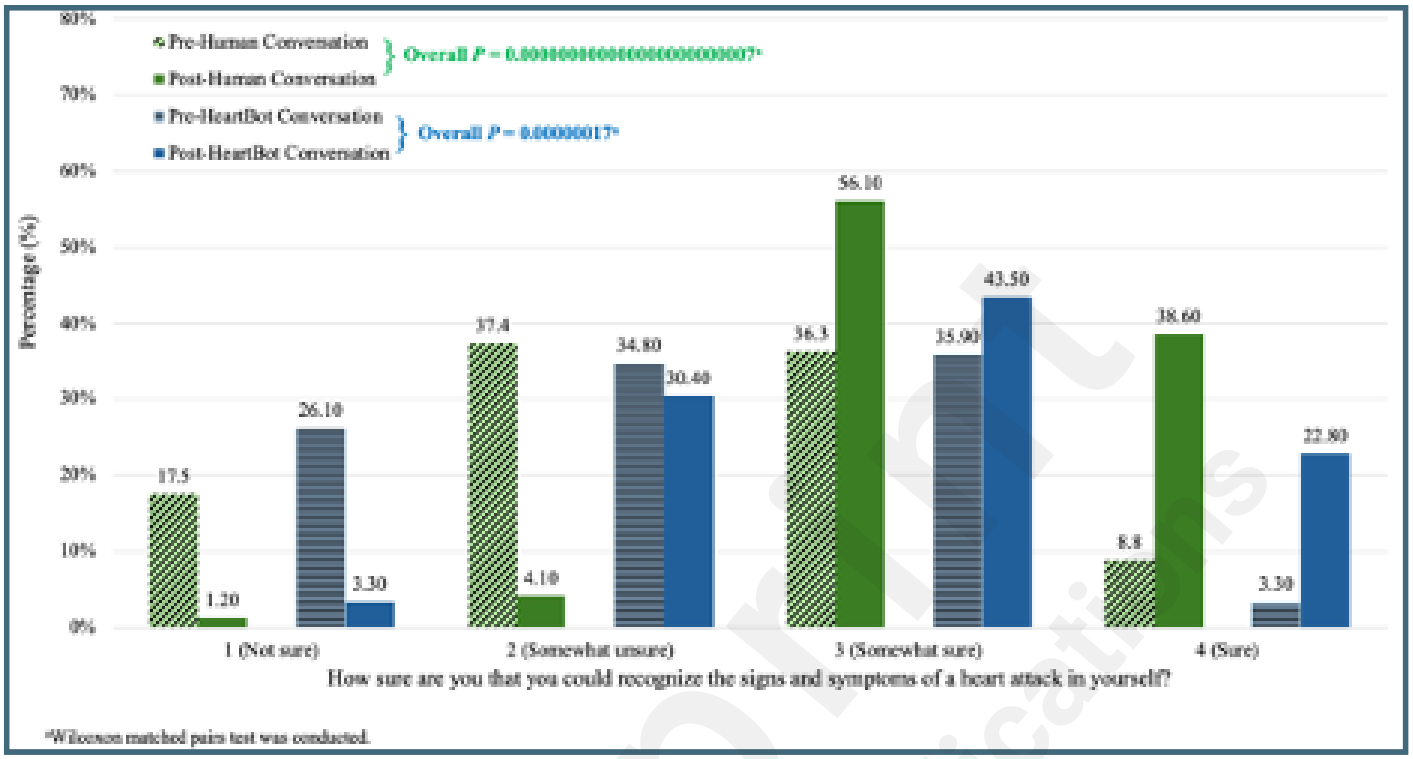
1. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. *Journal of Affective Disorders*. 2024;356:459-469. doi:10.1016/j.jad.2024.04.057
2. Kurniawan MH, Handiyani H, Nuraini T, Hariyati RTS, Sutrisno S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. Ann Med. 2024;56(1):2302980. doi:10.1080/07853890.2024.2302980
3. Oh, Y.J., Zhang, J., Fang, ML. et al. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 18, 160 (2021). https://doi.org/10.1186/s12966-021-01224-6
4. Noh E, Won J, Jo S, Hahm DH, Lee H. Conversational Agents for Body Weight Management: Systematic Review. J Med Internet Res. 2023;25:e42238. Published 2023 May 26. doi:10.2196/42238
5. Bendotti H, Lawler S, Chan GCK, Gartner C, Ireland D, Marshall HM. Conversational artificial intelligence interventions to support smoking cessation: A systematic review and meta-analysis. Digit Health. 2023;9:20552076231211634. Published 2023 Nov 3. doi:10.1177/20552076231211634
6. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc. 2018;25(9):1248–58.
7. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial Intelligence-Based Chatbots for Promoting Health Behavioral Changes: Systematic Review. J Med Internet Res. 2023;25:e40789. Published 2023 Feb 24. doi:10.2196/40789
8. Kim HK. The Effects of Artificial Intelligence Chatbots on Women's Health: A Systematic Review and Meta-Analysis. Healthcare (Basel). 2024;12(5):534. Published 2024 Feb 23. doi:10.3390/healthcare12050534
9. Lyzwinski LN, Elgendi M, Menon C. Conversational Agents and Avatars for Cardiometabolic Risk Factors and Lifestyle-Related Behaviors: Scoping Review. JMIR Mhealth Uhealth. 2023;11:e39649. Published 2023 May 25. doi:10.2196/39649
10. He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. Journal of medical Internet research, 25, e43862. https://doi-org.ucsf.idm.oclc.org/10.2196/43862
11. Lim SM, Shiau CW, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. Behavior Therapy. 2022 Mar 1;53(2):334-47. doi:10.1016/j.beth.2021.09.007
12. Terblanche N, Molyn J, de Haan E, Nilsson VO. Comparing artificial intelligence and human coaching goal attainment efficacy. Plos one. 2022 Jun 21;17(6):e0270255. Doi: 10.1371/journal.pone.0270255
13. Fukuoka Y, Kim D, Holli D, et al. Usability Testing of the HeartBot, an AI-Driven Conversational Agent, to Promote Women's Knowledge and Awareness of Heart Attack Risk and Response. JAMA Network. (in progress)
14. Go Red for Women. Homepage. American Heart Association. 2024. Available from: https://www.goredforwomen.org/en/
15. Dahlbäck N, Jönsson A, Ahrenberg L. Wizard of Oz studies: why and how. InProceedings of the 1st international conference on Intelligent user interfaces 1993 Feb 1 (pp. 193-200)
16. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet: Viewpoint. J Med Internet Res. 2020 Sep 30;22(9):e22845. doi: 10.2196/22845. PMID: 32996892; PMCID: PMC7557439.
17. Liao W, Oh YJ, Feng B, Zhang J. Understanding the Influence discrepancy between human and artificial agent in advice interactions: The role of stereotypical perception of agency. Communication Research. 2023;50(5):633-64. doi:10.1177/00936502221138427
18. Feng B. Testing an integrated model of advice giving in supportive interactions. Human Communication Research. 2009 Jan 1;35(1):115-29. doi:10.1111/j.1468- 2958.2008.01340.x
19. Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J of Soc Robotics. 2009 Jan;1:71-81. doi:10.1007/s12369-008-0001-`3
20. Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).
21. Martin SS, Aday AW, Almarzooq ZI, et al. 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. Circulation. 2024;149(8):e347-e913. doi:10.1161/CIR.0000000000001209
22. Harris PA, Taylor, R, Thielke, R, et al. Research electronic data capture (REDCap)– a metadata-driven methodology and workfow process for providing translational research informatics support. Journal of biomedical informatics. 2009;42(2):377-

381. doi:https://doi.org/10.1016/j.jbi
23. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/
24. Christensen RHB. ordinal - Regression Models for Ordinal Data. R package version 2022.11-16. Vienna, Austria: Comprehensive R Archive Network (CRAN); 2022. Available from: https://CRAN.R-project.org/package=ordinal
25. Stata Statistical Software: Release 16.1. StataCorp LLC; 2020.
26. Dergaa I, Ben Saad H, Glenn JM, Amamou B, Ben Aissa M, Guelmami N, Fekih-Romdhane F, Chamari K. From tools to threats: a reflection on the impact of artificial-intelligence chatbots on cognitive health. Frontiers in psychology. 2024 Apr 2;15:1259845.
27. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: Ethical issues with using chatbots in mental health. Digital health. 2023 Jun;9:20552076231183542.
28. Nass C, Steuer J, Tauber ER. Computers are social actors. InProceedings of the SIGCHI conference on Human factors in computing systems 1994 Apr 24 (pp. 72-78).
29. Bickmore TW, Picard RW. Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction (TOCHI). 2005 Jun 1;12(2):293-327.
30. Brave S, Nass C, Hutchinson K. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. International journal of human-computer studies. 2005 Feb 1;62(2):161-78.
31. Bickmore TW, Mitchell SE, Jack BW, Paasche-Orlow MK, Pfeifer LM, O'Donnell J. Response to a relational agent by hospital patients with depressive symptoms. Interacting with computers. 2010 Jul 1;22(4):289-98.

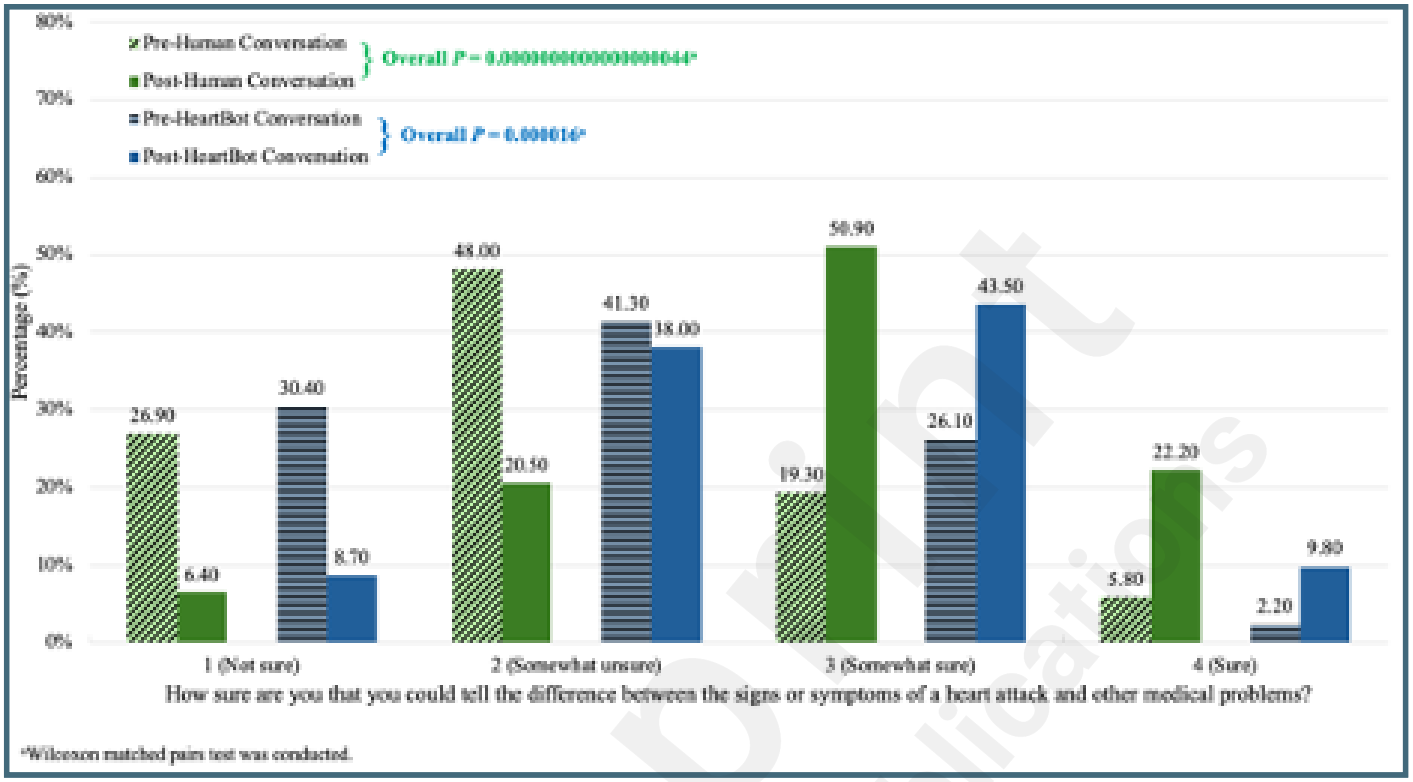
Supplementary Files

Figures

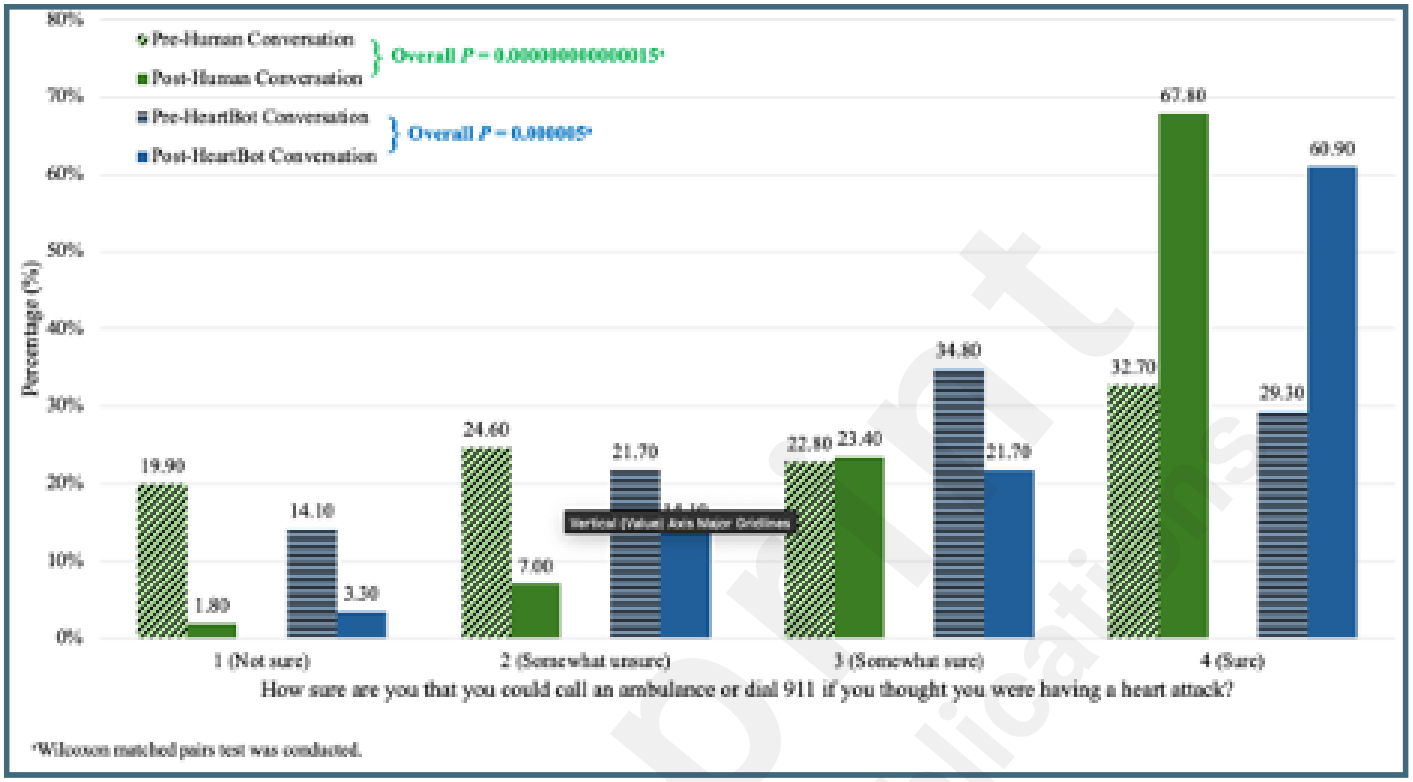
Change in response between pre and post Human conversation (n=171), and pre and post HeartBot conversation (n=92) for knowledge on recognizing signs and symptoms of a heart attack.



Change in response between pre and post Human conversation (n=171), and pre and post HeartBot conversation (n=92) for knowledge on telling the difference between the signs or symptoms of a heart attack and other medical problems.



Change in response between pre and post Human conversation (n=171), and pre and post HeartBot conversation (n=92) for knowledge on calling an ambulance or dialing 911 when experiencing heart attack.



Change in response between pre and post Human conversation (n=171), and pre and post HeartBot conversation (n=92) for knowledge on getting to an emergency room within 60 minutes after onset of symptoms of a heart attack.