JOURNAL OF
EMERGING INVESTIGATORS

Article

Trust in the use of artificial intelligence technology for treatment planning

Meenal Srivastava[1]*, Janvi Srivastava[1]*, Adamaris Regalado[2], Abril Augustin[3], Lisa Worthy[4]

[1] BASIS Scottsdale, Scottsdale, Arizona
[2] Arizona State University, Tempe, Arizona
[3] Peoria High School, Peoria, Arizona
[4] Psychology Department, Glendale Community College, Glendale, Arizona

* These authors contributed equally to this work.

SUMMARY

As artificial intelligence (AI) becomes more prevalent in day-to-day life, it is important to consider public opinion and acceptance towards these AI systems. Specifically, many struggle to trust AI when used to create medical treatment plans. After all, one's health tends to be a very emotionally-charged issue and not necessarily what we would associate with a machine. To address this, we present the question: Do young college students from diverse backgrounds trust AI system-developed treatment plans? We hypothesized that participants would rate the treatment plan developed by the AI system lower than the treatment plan developed by a physician. We conducted a between-group randomized controlled experiment with 81 community college students (75% female, 25% male) from a Hispanic Serving Institution. We presented the control group with a case study in which a physician designed the treatment plan. We presented the experimental group with a case study in which an AI system designed the treatment plan. The AI-developed treatment plan scored lower on the trust rating scale than the physician-created treatment plan, which is consistent with the hypothesis. There was no statistically significant difference between the two groups' scores on the Healthcare Trust Questionnaire. Our results also showed no significant difference between the trust levels in AI of people of different ages, genders, ethnicities, employment statuses, or hospitalization statuses, contradicting previous research. Overall, our findings may indicate a negative public opinion regarding AI-developed treatment plans, potentially deterring the future of AI-driven healthcare.

INTRODUCTION

Artificial Intelligence (AI) consists of computer systems that can perform tasks that usually require human intelligence–such as language translation, image recognition, and large-scale data analysis – and have been used in disciplines such as/including medicine, literature, and engineering (1). Previous research examining trust–defined as willingness to rely upon–and the use of AI in healthcare applied different types of research designs (e.g., between and within group, focus groups, case studies) across many different populations including patients, physicians, clinicians (1–3). Results on trust of AI have been mixed. A cross-cultural examination of trust in AI found that approximately 37% of respondents were unwilling to rely on information provided by AI and approximately 30% were unwilling to rely on AI for healthcare (1). Results have demonstrated that several factors appear to influence trust in AI including age, education, employment, prior hospitalization, severity of medical condition, confidence in physicians, and cultural background (1–3). Younger populations (e.g., Millennials and Gen Z) tend to be more trusting of AI and digital technology in healthcare compared to other generational cohorts (1,3,4). Females tend to view AI more negatively compared to males (5). Those with a college education were more likely to trust and approve of AI in general (1,3). Research has demonstrated that characteristics of the AI system, including ease of explanation and physical invasiveness of procedure (eg., dermatology versus robotic surgery) impact trust in AI (3,6).

Our review of the literature found no experimental design that used a case study, describing a medical condition that disproportionately impacts young adults (ages 16–24), administered to an ethnically diverse student population at a Hispanic Serving Institution (HSI). Our study adds to the literature by asking specifically about trust in AI treatment plan development, which represents a shift from prior research that has primarily examined AI use for clinical diagnosis and screening. The aim of this study is to determine whether our participants trust a treatment plan developed by human physicians more than an AI system-developed treatment plan. We hypothesized that if presented with a treatment plan, participants would rate the treatment plan developed by the AI system lower than the treatment plan developed by the physician. These ratings are based on how willing the participant would be to use the treatment plan. The most interesting result from our study was that there was a trend towards less trust in the AI-developed treatment plan. Our research speaks to the broader context of AI usage and trust, especially as we start to adopt AI in all areas of our lives–from medicine to education to customer service. We hope to provide useful information on young people's stance on medical AI for future policymakers.

	Mean Trust Rating in Treatment Plan	Mean Score in Healthcare Trust Questionnaire
Control (physician-developed treatment plan)	8.636	42.909
Experimental (AI-developed treatment plan)	6.875	43.021
p-value	<0.01***	0.954

Table 1: Trust ratings in treatment plans developed by a physician compared to an AI with the corresponding scores in Healthcare Trust Questionnaire. Left column showing mean trust rating in treatment plans ± SD (n=81). Participants rated their trust in the treatment plan they were presented with on a scale of 1 to 10, with the experimental group's treatment plan created by an AI system and the control group's treatment plan created by a human physician (independent samples *t*-test, $p < 0.01$, $d = 1.15$). Right column showing mean ± SD (n=81). Participants answered questions regarding their trust in the healthcare system and AI applications in improving healthcare by ranking their agreement in various statements on a scale of 1 to 6, from strongly disagree to strongly agree (*t*-test, $p = 0.954$).

RESULTS

Remember to try to place figures after its first inline reference. We aimed to determine which type of treatment plan our participants would be more willing to rely on – a treatment plan developed by human physicians or an AI system. We hypothesized that if presented with a treatment plan, participants would rate the treatment plan developed by the AI system lower than the treatment plan developed by the physician. In order to address this hypothesis, we randomly assigned participants to read identical passages describing a scenario in which a patient was diagnosed with aseptic meningitis and sent for further treatment. Both passages were identical. The only difference was that in the AI scenario, it was clearly written that the treatment plan was made by an AI model specializing in medical planning, whereas in the physician scenario, it was stated that the treatment plan was made by an experienced doctor. For more information on the demographic makeup of the participants, see appendix.

Participants who read the scenario about the AI-generated treatment plan rated their trust–based on the criteria of how willing they'd be to use the treatment plan if they had aseptic meningitis–significantly lower on the treatment planning process item than participants who read the scenario about the physician created treatment plan ($t_{(79)} = 5.09$, $p < 0.001$, $d = 1.15$). Cohen's d indicates a large effect of the applied treatment when > 0.8. (**Table 1**). Subsequent analyses found no statistically significant difference between the case studies and ratings of overall healthcare trust as measured by the Healthcare Trust Questionnaire ($t_{(79)} = -0.06$, $p = 0.95$; **Table 1**). Contrary to previous research by Gillespie et al. (1), we found no statistically significant difference in ethnicity and trust in AI in healthcare ($F_{(2, 77)} = 1.50$, $p = 0.23$), as indicated by ratings of trust in the treatment plans ($F_{(2, 77)} = 1.50$, $p = 0.23$; **Figure 1**) or the responses to the Healthcare Trust Questionnaire (**Figure 2**). An additional analysis found no statistically significant difference in gender and trust in AI in healthcare ($t_{(77)} = 1.833$, $p = 0.063$), which is inconsistent with previous research (3). Prior hospitalizations did not have a statistically significant impact on overall trust of technology and AI in healthcare as measured by the Healthcare Trust Questionnaire ($F_{(2, 78)} = 0.51$, $p = 0.61$; **Figure 3**) or a statistically significant impact on trust in the treatment planning process (*F(2, 78)* = 2.28, $p = 0.11$; **Figure 4**), which was also inconsistent with previous research (3). Generational differences were not examined because there were too few members for some categories (e.g., people over 65); however, correlation results for age and trust were not statistically significant (Pearson Correlation r = 0.342, $p = 0.30$).

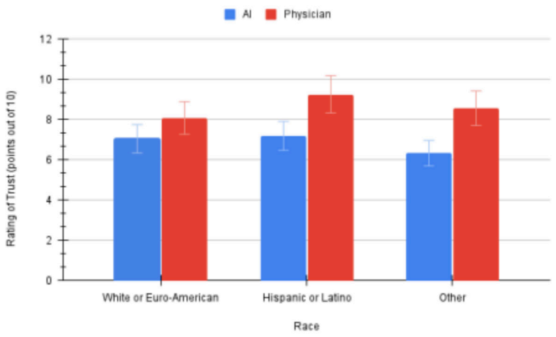

Figure 1: Trust ratings in treatment plans developed by a physician compared to an AI across racial groups. Bar graph showing mean trust rating ± SD (n=81). Participants rated their trust in the treatment plan they were presented with on a scale of 1 to 10, with the experimental group's treatment plan being made by an AI (blue) and the control group's treatment plan being made by a human physician (red). These scores were compared between three racial groups. There was no significant difference between ratings of trust for all three groups.

DISCUSSION

Participants in the experimental group, who read about an AI system-developed treatment plan for meningitis rated their trust significantly lower than participants who read about a physician-developed treatment plan for the same condition. There was no statistically significant difference between the scores on the Healthcare Trust Questionnaire,

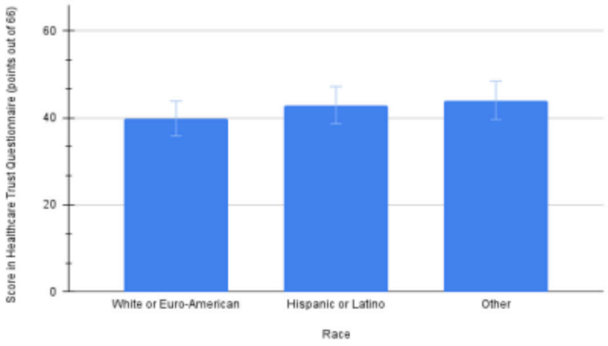

Figure 2: Score in the Healthcare Trust Questionnaire among racial groups. Bar graph showing mean score in Healthcare Trust Questionnaire ± SD (n=81). These scores were compared between 3 racial groups (White/Euro-American, Hispanic/Latino, or other). Participants answered questions regarding their trust in the healthcare system and AI's applications in improving healthcare by ranking their agreement in various statements on a scale of 1 to 6, from strongly disagree to strongly agree, with a maximum of 66 points possible. There was no significant difference between scores in the Healthcare Trust Questionnaire for all three groups.

https://doi.org/10.59720/23-289

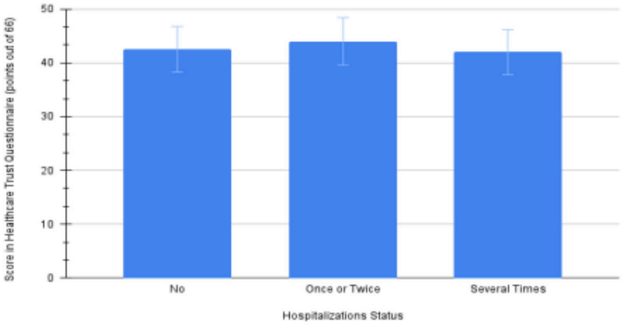

Figure 3: Trust ratings in treatment plans developed by a physician compared to an AI among participants of different hospitalization statuses. Bar graph showing mean (n=81). Participants rated their trust in the treatment plan they were presented with on a scale of 1 to 10, with the experimental group's treatment plan being made by an AI and the control group's treatment plan being made by a human physician. These scores were compared between three hospitalization statuses. There was no significant difference between rating of trust for an AI-created treatment plan and a physician-created treatment plan across hospitalization statuses.

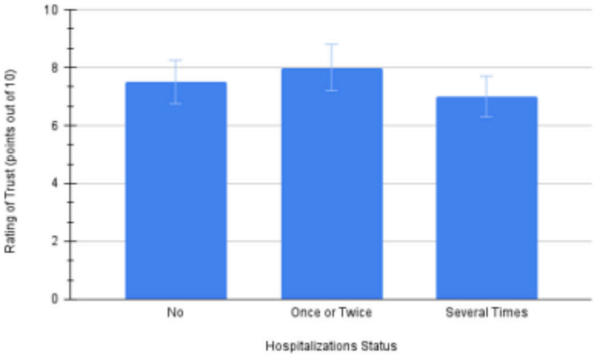

Figure 4: Score on the Healthcare Trust Questionnaire among participants of different hospitalization statuses. Bar graph showing mean score on the Healthcare Trust Questionnaire ± SE (n=81). These scores were compared between three hospitalization statuses (Zero, Once or Twice, Three or More). Participants answered questions regarding their trust in the healthcare system and AI's applications in improving healthcare by ranking their agreement in various statements on a scale of 1 to 6, from strongly disagree to strongly agree, with a maximum of 66 points possible.

which measured trust in AI and healthcare in general (to see questions asked, see appendix).

The discrepancy between ratings in trust in a case study about AI-generated treatment plan versus ratings of trust in a survey about AI's utility in healthcare in general may have occurred because the case study made individuals feel a personal connection to the case study, causing participants to reveal their true reactions (increased trust in humans over AI). In contrast, when asked broad questions about AI's utility in improving healthcare, participants may claim to trust AI because the question feels less applicable to themselves–the Healthcare Trust Questionnaire asked for participants' general opinions about AI in healthcare, whereas the case study asked them to put themselves in the shoes of a patient. This effect was seen when participants rated trust in AI lower when presented with a case study about an AI-generated treatment plan than when answering our Healthcare Trust Questionnaire's survey questions about AI's utility in healthcare planning.

Our results did not align with the literature in several ways. We found no statistically significant effects of age, prior hospitalization, gender, or ethnicity on scores on the Healthcare Trust Questionnaire or rating of trust in either the physician-developed treatment plan or the

AI-developed treatment plan, contradicting most prior research on this topic (1–3). It is difficult to uncover the underlying reason for this difference. We propose that our results were obtained due to the unique setting of our study: a HSI in the Southwestern US with limited racial diversity, thereby limiting the heterogeneity of opinions represented.

One key limitation of our study is the lack of diversity within the gender, ethnicity, age, education level, and career path of the sample. For instance, the participants consisted primarily of self-identified women (n=61), as compared to men (n=18). Furthermore, the ethnic diversity of the participants was limited, with over 72% of all participants being either White/Euro-American or Latino/Hispanic. Future research should consider broadening the sample to reflect differences in age, education, and career path. Lastly, different participants may have had different ideas of what healthcare technology is (e.g., X-rays or imaging instead of AI), which may have led to results in the Healthcare Trust Questionnaire that do not truly represent public trust in AI. Future research should ask participants what they consider to be technology in healthcare, because different schemas may result in different levels of trust.

According to our findings, participants tended to trust healthcare, technology, and medical AI in general, but tended to rate trust lower when an AI-system was used to develop a treatment plan for a condition more likely to impact young adults. Therefore, we propose that campaigns to improve trust of medical AI specifically address public perception of treatment plans made by AI, rather than medical AI in general, as in most current advertising models.

Finally, it is important that future policy makers do not continue to fall into the trap of assuming that young people

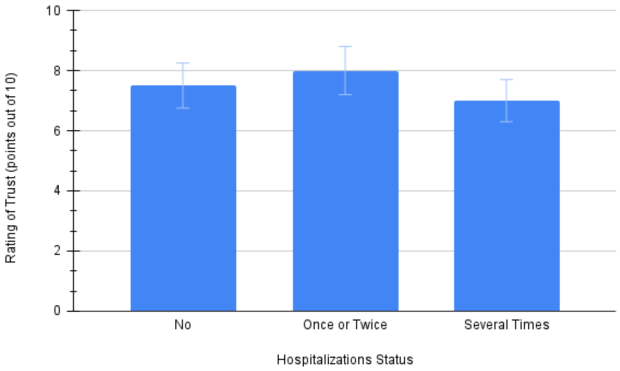

Figure 5: Scores on the Healthcare Trust Questionnaire across genders. Bar graph showing mean (n=81). These scores were compared between two genders. Participants answered questions regarding their trust in the healthcare system and AI's applications in improving healthcare by ranking their agreement in various statements on a scale of 1 to 6, from strongly disagree to strongly agree.

https://doi.org/10.59720/23-289

trust technologies to make decisions for them, as this does not seem to be the case specifically for AI-system developed treatment plans. If AI is continued to be used for the creation of treatment plans, we may see shorter wait times for appointments and reduced employment of healthcare workers. We must also consider that AI models that have been trained on largely homogeneous populations can lead to worse healthcare outcomes for minority groups–thereby exacerbating the inequalities in medical access we already see today between people of minority races, gender identities, etc.

MATERIALS AND METHODS

The study used a between-group randomized controlled experimental design with undergraduate students at an HSI, who were enrolled in summer psychology courses. Participants were recruited using a block randomization, clustered convenience sampling technique. The sample consisted of 81 undergraduates (61 women, 18 men, 2 non-binary/did not specify–other gender identities were excluded due to an extremely small sample), ranging in age from 17–56 years with a mean of 24.71 years. The sample was primarily White (38.27%) and Hispanic/Latino (34.57%). One set of participant data was excluded because the individual completed the experiment twice. There were 33 participants in the control condition and 48 participants in the experimental condition. Ethical standards were met, and the project received IRB approval.

Faculty in the psychology department who were teaching a summer course were contacted by the researchers and invited to participate in the study. Courses taught by willing faculty were randomly assigned to either the experimental group or control group. Faculty then shared either a link to the control condition (i.e., physician created treatment plan, **Figure 1**) or a link to the experimental condition (i.e., AI – system created treatment plan, **Figure 1**) with their students.

We chose a between-group design as opposed to a within-group design because if participants were exposed to both conditions, as with a within-group experiment, the order in which they read the control or experimental case studies may have influenced their responses. It may have also led to participants going back and changing their survey responses to a case study after reading the second one. Hence, we determined that ensuring each participant was only exposed to one case study would ensure more reliable survey data.

The study took about 15–20 minutes to complete. Participants were not masked to condition assignment; however, they were unaware that there was another version of the survey they were taking–hence keeping them unaware that there was both a control group and an experimental group. Investigators and those assessing outcomes were aware that there were two groups and which group each participant was assigned to. Participants who consented to participate read a scenario and completed two rating scales assessing trust in the treatment plan, healthcare, and technology in healthcare. The independent variable was the method of treatment plan development including two levels: the AI system-developed treatment plan, and the physician-developed treatment plan. Participants in both groups (control and experimental) were asked to read a scenario that described meningitis, a health condition that impacts young adults more than the general population. The experimental condition scenario indicated that an AI system created a treatment plan to address meningitis. The control condition scenario indicated that a highly qualified physician created a treatment plan to address meningitis. The researchers' first dependent variable was a 10-point rating scale of trust in the treatment plan from 1 being *no trust* to 10 being *complete trust*.

The second dependent variable was a measure of trust in healthcare, healthcare providers, and use of technology in healthcare. Level of trust in healthcare was assessed by the Healthcare Trust Questionnaire with a six-point rating for each question from *strongly disagree* to *strongly agree*. Questions focused on participants' opinions on the utility of both AI and medical AI in improving healthcare access and accuracy. Examples include, "*I trust technology, like artificial intelligence, to provide comprehensive treatment plans after a condition has been diagnosed*" and "*I believe that technology used in healthcare, like artificial intelligence, will lead to improved quality of life.*"

This Healthcare Trust Questionnaire was created using items identified in previously published research (1). Eleven items were combined into a single overall score. The internal consistency of the Healthcare Trust Questionnaire using Cronbach's alpha was 0.81, which is considered good consistency for a social science measure (5).

Independent samples t-tests were used for analyses involving one categorical variable across two groups. We used these to compare the results of the Healthcare Trust Questionnaire between the control group and experimental group and to compare ratings of trust in the treatment plans between the control group and experimental group. This was done to identify significant differences between groups.

A one-way linear ANOVA was used for analyses involving one categorical variable across three or more levels. We used these to compare the results of the Healthcare Trust Questionnaire between participants with 0 recent hospitalizations, 1–2 recent hospitalizations, and several recent hospitalizations. This was also used to compare ratings of trust in the treatment plans between participants with 0 recent hospitalizations, 1–2 recent hospitalizations, and several recent hospitalizations. This was done to identify significant differences between groups. The assumptions of homogeneity of variances and normality of data were met.

A two-way linear ANOVA was used for analyses involving two categorical variables across two to three levels. We used these to compare ratings of trust in the treatment plans between the control group and experimental group across three racial groups: White/Euro-American participants, Hispanic participants, and participants identifying with another race. This was done to identify significant differences between groups. The assumptions of homogeneity of variances and normality of data were met.

Received: September 27, 2023
Accepted: April 24, 2023
Published: September 18, 2024

REFERENCES

1. Gillespie, Nicole, et al. “Trust in Artificial Intelligence: A Five Country Study.” *The University of Queensland; KPMG*, March 2021, https://doi.org/10.14264/e34bfa3.
2. Ploug, Thomas, et al. “Population Preferences for Performance and Explainability of Artificial Intelligence

in Health Care: Choice-Based Conjoint Survey." *Journal of Medical Internet Research*, vol. 23, no. 12, Dec. 2021, https://doi.org/10.2196/26611.

3. Yakar, Derya, et al. "Do People Favor Artificial Intelligence over Physicians? A Survey among the General Population and Their View on Artificial Intelligence in Medicine." *Value in Health*, vol. 25, no. 3, Oct. 2021, pp. 374–381., https://doi.org/10.1016/j.jval.2021.09.004.
4. Alkire (née Nasr), Linda, et al. "Patient Experience in the Digital Age: An Investigation into the Effect of Generational Cohorts." *Journal of Retailing and Consumer Services*, vol. 57, Nov. 2020, p. 102221., https://doi.org/10.1016/j.jretconser.2020.102221.
5. Novozhilova, Ekaterina, et al. "More Capable, Less Benevolent: Trust Perceptions of AI Systems across Societal Contexts." *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, Feb. 2024, pp. 342-366, https://doi.org/10.3390/make6010017.
6. George, Darren, and Paul Mallery. *IBM SPSS statistics 19 step by step*. Pearson Education. Jan. 2013.
7. Harmon-Jones, Eddie, and Judson Mills. "An Introduction to Cognitive Dissonance Theory and an Overview of Current Perspectives on the Theory." *American Psychological Association*, 2019, www.apa.org/pubs/books/Cognitive-Dissonance-Intro Sample.pdf.

Appendix

Control scenario. The scenario presented to the control condition participants describes a young adult diagnosed with meningitis who was referred to an experienced physician for treatment planning. Meningitis is the swelling of the protective membranes covering the brain and spinal cord. It can be caused by either bacterial or viral infection of the fluid surrounding the brain or spinal cord. Young adults (16–23) have an increased risk of contracting meningitis, as compared to the general population. Assume that a 20-year-old patient presents to the emergency room complaining of fever, a headache, and a stiff neck. The patient is otherwise previously healthy. After examining the patient further, the patient was diagnosed with aseptic meningitis–inflammation of the brain meninges due to a reason other than bacterial infection–and was referred for further treatment planning. A comprehensive treatment plan that included all necessary medical interventions, including selecting the appropriate medication, dosage of the medication, and follow-up care for the patient in the scenario was created by a physician. The physician is a highly qualified healthcare provider with 25 years of practicing neurological care with a specialization in conditions that impact young adults.

Experimental scenario. The scenario presented to the experimental condition participants describes a young adult diagnosed with meningitis who was referred to an artificial intelligence system for treatment planning. Meningitis is the swelling of the protective membranes covering the brain and spinal cord. It can be caused by either bacterial or viral infection of the fluid surrounding the brain or spinal cord. Young adults (16–23) have an increased risk of contracting meningitis, as compared to the general population. Assume that a 20-year-old patient presents to the emergency room complaining of fever, a headache, and a stiff neck. The patient is otherwise previously healthy. After further examination, the patient was diagnosed with aseptic meningitis and was referred for further treatment planning. An artificial intelligence (AI) system was used to create a comprehensive treatment plan to provide the necessary medical interventions, including selecting the appropriate medication, dosage of the medication, and follow-up care for the patient in the scenario. Artificial intelligence in healthcare is a broad term used to describe the use of machine learning algorithms and software, or artificial intelligence (AI), to mimic human thinking in the analysis, presentation, and comprehension of complex medical and health care data. The AI system was built by experts in the field of artificial intelligence, trained using 2 million diagnostic contributions from patients and tested by physicians. The AI system training included conditions that impact young adults.

Healthcare Trust Questionnaire presented to all participants:

Items	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
I have trust in the healthcare system.						
I have trust in physicians and other healthcare providers to create an effective treatment plan for me.						
I have trust in technology used in healthcare to create an effective treatment plan for me.						
I trust physicians to provide comprehensive treatment plans after a condition has been diagnosed.						
I trust technology, like artificial intelligence, to provide comprehensive treatment plans after a condition has been diagnosed.						
I believe that technology used in health care, like artificial intelligence, will lead to improved quality of life.						
I believe that technology used in health care, like						

artificial intelligence, will lead to peace and political stability.						
Even if computers are better at evaluating medical conditions, I still prefer a doctor.						
I think medicine is not ready for implementing artificial intelligence in decision making, including creating treatment plans.						
Through human experience a physician or health care provider can detect more than a computer.						
I do not have trust in the healthcare system.						

Demographic Makeup of Participants

	Control (Physician)	**Experimental Group (AI)**
Gender		
Man	12 (36.36%)	6 (12.50%)
Woman	20 (60.61%)	41 (85.42%)
Non-Binary	0 (0%)	1 (2.08%)
Other	1 (3.03%)	0 (0%)
Ethnicity		

American Indian or Alaskan Native	2 (6.06%)	1 (2.08%)
White or Euro-American	12 (36.36%)	19 (39.58%)
Asian or Asian-American	1 (3.03%)	3 (6.25%)
Native Hawaiian or other Pacific Islander	0 (0%)	0 (0%)
Black or African American	2 (6.06%)	4 (8.33%)
Latino or Hispanic	12 (36.36%)	16 (33.33%)
Other	4 (12.12%)	5 (10.42%)
Frequency of Medical Treatment in the Last Year		
Several Times	8 (24.24%)	12 (25%)
1-2 Times	15 (45.45%)	17 (35.42%)
Never	10 (30.30%)	19 (39.58%)
Highest Education Level		
Less than High School Graduate	1 (3.03%)	0 (0%)
High School Graduate	30 (90.91%)	32 (66.67%)
College Graduate	2 (6.06%)	16 (33.33%)
Age	24	25