Using ChatGPT 4 as an Aid in Medical Diagnosis and Treatment: Assessing Gender Bias

Keira M. Cornwell

1 Introduction

Gender bias in medicine is a pressing issue. While 44% of women in the U.S. are living with heart disease, women are twice as likely to die of heart disease than men due to doctors misunderstanding and misdiagnosing heart disease symptoms in women [1]. As more doctors begin to use LLMs in the clinic, and as ones like ChatGPT 4 begin to outperform doctors [2], it is critical to understand if LLMs exhibit a similar medical bias to doctors. In this paper, I provide an overview of what I found through the literature, as well as my own experimentation, in an effort to better understand gender bias from LLMs and inform the future role of chatbots in the clinic.

2 Methods

2.1 Assessing Difference in Diagnostic Tone and Urgency

I began to identify metrics for gender bias in medical diagnosis by conducting a literature review of how conditions are commonly misdiagnosed, delayed in diagnosis, or missed altogether for women in comparison to men. In their 2021 paper, Sun et. al. identify 112 acute and chronic disease phenotypes that are not sex-specific and would have the same underlying diagnosis. Sun et. al. found by analyzing over 157 million patient records from Medicaid, Medicare, and Columbia University Irving Medical Center, that women experienced a statistically significant ($p < 0.01$) diagnostic delay for almost all of these conditions, for a majority of the conditions, finding most prominently that they experienced a delay for 98 of the 112 conditions for the Medicaid dataset [3].

While there was not a direct test I could run using ChatGPT 4 to see if there was a diagnostic delay between male and female patients, the findings of Sun et. al. inspired me to analyze if ChatGPT 4 would analyze the same set of symptoms, pointing to the same underlying diagnosis, for male and female patients similarly. Specifically, I was interested in testing to see if they treated both patients with a similar tone and level of urgency, and if similar potential causes were identified. I pursued these research questions through two methods, which I will proceed to discuss.

2.1.1 Data Collection

To collect data on Chat GPT's responses to different sex patients with the same symptoms, I began by assembling a list of 96 common symptoms identified from NHS Inform [4] associated with the acute and chronic conditions that Sun et. al. found were associated with diagnostic delay for women, which ranged in severity from fatigue to shooting pain in the arms. For 500 trials, I randomly chose one to seven symptoms from the list, and prompted ChatGPT 4 with the following template: "I am a *{gender}* experiencing the following symptoms: *{symptoms}*. What might be causing this, and what should I do?" For each prompt, I used the same symptoms for both one female patient and one female patient, ending with a total list of 1000 ChatGPT 4 responses.

2.1.2 Bootstrapping Responses

To assess if there was a statistically different tone and response type between the male and female sex, I conducted a bootstrapping analysis. I embedded each response using the OpenAI text-embedding-3-large model, which measures the relatedness between two pieces of text. I separated the responses between the female and male classifiers, and calculated the mean cosine similarity, using the sklearn toolkit, of each male/female response pair for the same prompt. I used cosine similarity as opposed to Euclidian, as it is a better way to calculate the similarity between vectors for high-dimensional vectors, such as text embeddings. I then performed a bootstrapping analysis to simulate an entire population by repeating this process for a random two groups of 500 prompts each for 10,000 trials, and finally returned the p-value.

2.1.3 Training a BERT Model

I was then interested to see if the responses were different enough that I could train a model to predict if the response was for a male or female based on the given response. The model I chose to train was a Bidirectional Encoder Representations from Transformers (BERT) model, which represents input text in a high-dimensional space, which allows to understand the context of words in nuanced sentences, making it a state-of-the-art natural language processing model. To create and train my BERT model for sex-classification, I followed Medium's "Text Classification with BERT" tutorial, and adapted the code for my data set [5]. Before entering a prompt into my model, I replaced any mention of "female," "male," "man," and "woman" in the prompt with "person." None of the responses contained the use of pronouns.

2.2 Assessing Difference in Pain Management

Upon conducting my literature review, I found a 2021 paper written by six Stanford students who were interested in a similar question of medical gender bias from ChatGPT. In particular, they focused on the fact that women are less likely to be prescribed pain medication for pain management than men. The students partnered with physicians to develop 55 clinical vignettes, where they prompted ChatGPT 2 with a theoretical pain management for five different types conditions: acute cancerous condition, chronic cancerous condition, acute non-cancerous condition, chronic non-cancerous condition, and post-op care. They prompted ChatGPT with each prompt for different combinations of race and gender, doing each combination once because they used a temperature of 0, and received the log-likelihood from each response. For each vignette, they gave ChatGPT a theoretical painkiller and asked if they would prescribe it or not, and if so, at a high or low dosage.

Unfortunately, OpenAI no longer makes the log likelihood of its responses publicly available, so I replicated their experiment differently. First, to limit the scope of my research to more directly relate to the first phase of my project, I only varied gender and not race. Second, since I did not have access to log-likelihoods of responses, I instead used a temperature of 0.7, and gave it the same vignette for each trial 11 times. Instead of finding the log-likelihood, I used a Wilcoxon rank sum test to find if there was a statistically significant difference between the probability of women not receiving a painkiller prescription in comparison to men and the probability of women receiving a low dosage in comparison to men. I chose a Wilcoxon rank sum test because the paper used paired t-tests, which I could not use since I did not have normally distributed data and was instead dealing with binary values (yes/no and high/low). However, a Wilcoxon rank sum test is described as the "non-parametric version of the two-sample t-test," which made it a comparable statistic metric for

my replication of their experiment [6]. However, I did use the same clinically developed vignettes as provided in the paper for my trials.

2.2.1 Wilcoxon Rank Sum

To perform the Wilcoxon rank sum test, I used the wilcoxon function in the scipy library. However, conceptually, the Wilcoxon rank sum test works by for each of the vignettes, summing the amount of times women were told "no" for a prescription and then doing the same for men. The absolute value of the difference between these sums is then calculated, and after taking out differences of 0, the differences are "ranked" such that the smallest difference has a rank of 1, up to the total data set size, n. Depending on if the difference, without the absolute value, is positive or negative, the symbols is then added back, and the wilcoxon test statistic and p-value is obtained.

3 Results

3.1 Bootstrapping Test

With bootstrapping the embedded form of the responses, the p-value was found to equal 0.484. This p-value is consistent with the fact that there was an equal 50/50 split in the data between men and women classifiers, and demonstrates that the cosine similarity calculated for the men and women stratified grouping was only more extreme than the cosine similarity for a random sample around 50% of the time. This p-value is thus not statistically significant, and furthermore, signifies that there was not a significant difference found in the response for men and women.

4.2 BERT Model Training

Using 4 epochs, or training rounds, the model ultimately had a prediction accuracy of 54%. The minimum accuracy, around 49%, was found during the first epoch, where the model learned to just guess "female" each time, as that automatically gave the model a 50% success rate due to the equal distribution of sex. While the model began to learn in the subsequent epochs, beginning to predict "male," it did not greatly increase the accuracy of the model, indicating the BERT accuracy was not really able to detect differences in language between male and female responses.

4.3 Wilcoxon Rank Sum for Pain Management

After performing the Wilcoxon rank sum test on the eleven trials, it found a p-value of a no-response of 0.1025 and p-value for a low dosage of 0.8049. For the no-response p-value, it found that women were more likely to receive a no-response, with a man not receiving a single no response across any of the vignettes or trials. While a p-value of 0.1025 is not less than 0.05, and thus not statistically significant, it still does indicate a pattern of women being more likely to not receive a painkiller prescription from ChatGPT 4. Conversely, a p-value for a low dosage being 0.8049 indicates that there is not a statistically significant difference in the frequency with which women were given a low dosage of painkillers in comparison to men.

5 Conclusion

I aimed to assess gender bias in medical diagnosis and recommendations from ChatGPT 4. Overall, I found that ChatGPT 4 remains relatively unbiased, especially from my bootstrapping and BERT

model data. Conversely, I did find that GPT 4 exhibits some bias in regards to painkiller prescription recommendations for women in comparison to men, but only in the question of receiving a prescription, and not in regards to dosage. I find these results to be extremely promising, as as more doctors begin to use LLMs like GPT 4 in the clinic to help diagnose and treat patients, it's important that these LLMs do not have biases on the basis of identities, such as gender. For future research, building upon the Stanford students' 2021 paper, a similar experiment could be repeated for race and intersectional identities to further test for bias across other relevant patient identifiers.

References

[1] CDC. (2024). About Women and Heart Disease. https://www.cdc.gov/heart-disease/about/women-and-heart-disease.html.

[2] Shmerling, Robert H. (2024). Can AI Answer Medical Questions Better than Your Doctor?. https://www.health.harvard.edu/blog/can-ai-answer-medical-questions-better-than-your-doctor-202403273028.

[3] Sun, Tony Y., et. al (2023). Large-scale Characterization of Gender Differences in Diagnosis Prevalence and Time to Diagnose. Med Rx. https://pmc.ncbi.nlm.nih.gov/articles/PMC10592987/#ref-list1.

[4] NHS Inform (2025). Illnesses and Conditions. https://www.nhsinform.scot/illnesses-and-conditions/a-to-z/.

[5] Kang, Pham. (2023). Text Classification with BERT. https://medium.com/@khang.pham.exxact/text-classification-with-bert-7afaacc5e49b.

[6] Logê, Cécile, et. al. (2021). Q-Pain: A Question Answering Dataset to Measure Social Bias in Pain Management. https://arxiv.org/abs/2108.01764.

Use of AI

I used ChatGPT for this project. Specifically, I used it for idea generation (for a project that lied at the intersection of healthcare and gender, and different probability and statistics topics that may be relevant for my project). Additionally, I helped it debug and alter existing code to better fit my data. I also had it alter the provided vignettes to remove the race variable from the provided data. Lastly, I used to help me write the Wilcoxon rank sum function and generate a statistic from it that would be interpretable. I also used it to generate the code to tokenize the answers to the pain prompts for the wilcoxon test.

Acknowledgements

I discussed how to improve my project with both Anna Mattinger and Jerry Cain at office hours.