46. INVESTIGATING POTENTIAL GENDER DIFFERENCES IN CHATGPT-DIAGNOSED CLINICAL VIGNETTES

Anjali Mediboina[1], Meghana Bhupathi[2], Keerthana Janapareddy[3]

[1] MBBS, Intern, Alluri Sitarama Raju Academy of Medical Sciences, Eluru, India

[2] MBBS, Tutor, Alluri Sitarama Raju Academy of Medical Sciences, Eluru, India

[3] Medical Student, Gayatri Vidya Parishad Institute of Health Care and Medical Technology

BACKGROUND: The integration of artificial intelligence (AI) in medical decision-making introduces additional concerns, particularly regarding information bias within AI models such as ChatGPT, which heavily rely on training data. With gender-based disparities in diagnosis and treatment being well-documented in healthcare, there is a pressing need to evaluate the potential of AI models to perpetuate or alleviate these gender biases. **AIMS:** This study seeks to investigate gender differences in diagnostic accuracy within ChatGPT 3.5 by evaluating the accuracy and completeness of its responses to various clinical vignettes. **METHODS**: Ten medical conditions (including psychiatric, respiratory, cardiac, and cerebrovascular cases) previously reported for gender-based misdiagnoses, were selected for the study. Two identical clinical vignettes were created for each condition, with the only difference being the gender of the patient. These 20 vignettes were entered into ChatGPT 3.5 randomly by a single researcher, each accompanied by a prompt requesting the most likely explanation for the patient's symptoms and the next appropriate step in management. The responses generated by ChatGPT were evaluated for accuracy and completeness by two independent evaluators, utilizing a scale set by Johnson et al., which included a six-point Likert scale ranging from 1 (completely incorrect) to 6 (correct) for accuracy, and a three-point scale for completeness, ranging from 1 (incomplete) to 3 (comprehensive). Discrepancies were resolved through a blind consensus process. Data analysis and visualization was done using RStudio v4.3.2, with statistical significance between accuracy and completeness was determined using Spearman's R and Mann-Whitney U Tests. **RESULTS**: Among the 20 cases, six were incorrectly diagnosed, with two instances attributed to gender-based misdiagnoses. Specifically, ChatGPT misclassified ectopic pregnancy as appendicitis, and paroxysmal supraventricular tachycardia (PSVT) as a panic attack in female patients, despite indicative symptoms and prior correct diagnoses in male counterparts. Additionally, systemic lupus erythematosus (SLE) was inaccurately labeled as rheumatoid arthritis (RA) in both male and female patients. Moreover, eating disorders were misidentified, with ChatGPT failing to provide definitive diagnoses for these conditions. The overall median accuracy score was 6, (Mean = 5.5, SD = 0.6), while the median completeness score was 2.5 (Mean = 2.5, SD = 0.5). Correlation analysis indicated a non-significant relationship between accuracy and completeness (Spearman's R: rs = 0.23139, p = 0.3263), although Mann Whitney U test results suggested significant discrepancies in accuracy between correctly and incorrectly diagnosed cases (z-score = 5.39649, p < .00001). **CONCLUSION**: While the AI's responses were generally accurate and complete, the observed misdiagnoses of conditions such as PSVT and eating disorders highlight the need for a more thorough examination of potential biases in AI-driven chatbots. The varying outcomes in the Spearman's R and Mann-Whitney U tests indicate that, although there may not be a consistent linear relationship between accuracy and completeness, ChatGPT's performance differs significantly across scenarios, necessitating further investigation. Moreover, the small sample size of vignette may not fully capture the extent of potential biases. Despite these limitations, the findings underscore the complexity of AI in healthcare and the critical importance of continuous scrutiny and refinement of these models.

Table: Summary of Diagnostic Accuracy and Completeness of ChatGPT 3.5 for Gender-Based Clinical Vignettes.

Case Number*	Correct Diagnosis	Diagnosis by ChatGPT	Accuracy Score	Completeness Score
Case 1	ADHD	ADHD	6	3
Case 2	ADHD	ADHD	5	3
Case 3	Autism Spectrum Disorder	Autism Spectrum Disorder	5	3
Case 4	Autism Spectrum Disorder	Autism Spectrum Disorder	5	3
Case 5	Ectopic Pregnancy	Appendicitis	6	2
Case 6	Appendicitis	Appendicitis	6	2
Case 7	Multiple Sclerosis	Multiple Sclerosis	6	2
Case 8	Multiple Sclerosis	Multiple Sclerosis	6	2
Case 9	Chronic Obstructive Pulmonary Disease (COPD)	Chronic Obstructive Pulmonary Disease (COPD)	6	3
Case 10	Chronic Obstructive Pulmonary Disease (COPD)	Chronic Obstructive Pulmonary Disease (COPD)	6	3
Case 11	Asthma	Asthma	6	3
Case 12	Asthma	Asthma	6	2
Case 13	Transient Ischemic Attack (TIA)	Transient Ischemic Attack (TIA)	6	3
Case 14	Transient Ischemic Attack (TIA)	Transient Ischemic Attack (TIA)	6	3
Case 15	Paroxysmal Supraventricular Tachycardia (PSVT)	Panic Attack	5	2
Case 16	Paroxysmal Supraventricular Tachycardia (PSVT)	Paroxysmal Supraventricular Tachycardia (PSVT)	6	3
Case 17	Systemic Lupus Erythematosus (SLE)	Rheumatoid Arthritis	5	2
Case 18	Systemic Lupus Erythematosus (SLE)	Rheumatoid Arthritis	5	2
Case 19	Eating Disorder	No definitive diagnosis, concluded that the symptoms were due to a "combination of factors included hormonal imbalances, nutritional deficiencies, overtraining syndrome and psychological factors"	4	2
Case 20	Eating Disorder	Relative Energy Deficiency in Sport (RED-S)	5	2
Mean Score (SD)			5.55 (0.6)	2.5 (0.5)
Median Score			6	2.5
Spearman's R			rs = 0.23139, p = 0.3263.	
Mann-Whitney U			z-score = 5.39649, p < 0.00001	

Legend: *Case numbers 2, 4, 5, 7, 10, 11, 14, 15, 18, 20 were clinical vignettes of female patients. Case numbers 1, 3, 6, 8, 9, 12, 13, 16, 17, 19 were clinical vignettes of male patients.

Key Words: Artificial intelligence, ChatGPT, Gender bias, Healthcare, Misdiagnosis.