Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study

Travis Zack, Eric Lehman*, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, Atul J Butte, Emily Alsentzer*

Summary

Background Large language models (LLMs) such as GPT-4 hold great promise as transformative tools in health care, ranging from automating administrative tasks to augmenting clinical decision making. However, these models also pose a danger of perpetuating biases and delivering incorrect medical diagnoses, which can have a direct, harmful impact on medical care. We aimed to assess whether GPT-4 encodes racial and gender biases that impact its use in health care.

Methods Using the Azure OpenAI application interface, this model evaluation study tested whether GPT-4 encodes racial and gender biases and examined the impact of such biases on four potential applications of LLMs in the clinical domain—namely, medical education, diagnostic reasoning, clinical plan generation, and subjective patient assessment. We conducted experiments with prompts designed to resemble typical use of GPT-4 within clinical and medical education applications. We used clinical vignettes from NEJM Healer and from published research on implicit bias in health care. GPT-4 estimates of the demographic distribution of medical conditions were compared with true US prevalence estimates. Differential diagnosis and treatment planning were evaluated across demographic groups using standard statistical tests for significance between groups.

Findings We found that GPT-4 did not appropriately model the demographic diversity of medical conditions, consistently producing clinical vignettes that stereotype demographic presentations. The differential diagnoses created by GPT-4 for standardised clinical vignettes were more likely to include diagnoses that stereotype certain races, ethnicities, and genders. Assessment and plans created by the model showed significant association between demographic attributes and recommendations for more expensive procedures as well as differences in patient perception.

Interpretation Our findings highlight the urgent need for comprehensive and transparent bias assessments of LLM tools such as GPT-4 for intended use cases before they are integrated into clinical care. We discuss the potential sources of these biases and potential mitigation strategies before clinical implementation.

Funding Priscilla Chan and Mark Zuckerberg.

Lancet Digit Health 2024; 6: e12–22

See **Comment** page e2

*Equal contribution

Bakar Computational Health Sciences Institute (T Zack PhD, Prof A J Butte MD) **and Helen Diller Family Comprehensive Cancer Center** (T Zack), **University of California San Francisco, San Francisco, CA, USA; Computer Science and Artificial Intelligence Laboratory** (E Lehman MSc, Prof P Szolovits PhD) **and Laboratory for Computational Physiology** (Prof L A Celi MD), **Massachusetts Institute of Technology, Cambridge, MA, USA; Department of Computer Science** (M Suzgun, Prof D Jurafsky PhD), **Stanford Law School** (M Suzgun), **and Department of Linguistics** (Prof D Jurafsky), **Stanford University, Stanford, CA, USA; Division of General Internal Medicine** (J A Rodriguez MD, Prof D W Bates MD, E Alsentzer PhD) **and Division of Pulmonary and Critical Care Medicine** (Prof R-E E Abdulnour MD), **Brigham and Women's Hospital, Boston, MA, USA; Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA** (Prof L A Celi)**; Department of Biostatistics** (Prof L A Celi) **and Department of Health Policy and Management** (Prof D W Bates), **Harvard T H Chan School of Public Health, Boston, MA, USA; Department of Radiology, Emory University, Atlanta, GA, USA** (Prof J Gichoya MD)**; Harvard Medical School, Boston, MA, USA** (Prof R-E E Abdulnour, E Alsentzer)**; Center for Data-Driven Insights and Innovation, University of California, Office of the President, Oakland, CA, USA** (Prof A J Butte)

Introduction

Large language models (LLMs), such as ChatGPT[1] and GPT-4,[2] have shown immense promise for transforming health-care delivery and are rapidly being integrated into clinical practice.[3] Indeed, several LLM-based pilot programmes are underway in hospitals,[4] and clinicians have begun using ChatGPT to communicate with patients and draft clinical notes.[5] While LLM-based tools are being rapidly developed to automate administrative or documentation tasks, many clinicians also envision using LLMs for clinical decision support.[5–8]

LLM-based tools have shown great potential, but there is also cause for concern about using LLMs for clinical applications. Extensive research has shown the potential for language models to encode and perpetuate societal biases.[9–13] Language models are typically trained using vast corpora of human-generated text to predict subsequent text on the basis of the preceding words. Through this process, models can learn to perpetuate harmful biases seen in the training data.[14] Although some of these biases, once identified, can be addressed via additional targeted training through a process called reinforcement learning with human feedback, this is a human driven process, which can be imperfect and even introduce its own biases.[15–17] Encoded biases can lead to poorer performance for historically marginalised or under-represented groups. For example, in a recent study that leveraged an LLM trained on clinical notes for clinical and operational tasks, predictions of 30 day readmission were significantly worse for Black patients than for other demographic groups.[18]

Our objective was to measure the propensity of GPT-4 to encode racial and gender biases and examine potential harms that might result from the use of GPT-4 in clinical applications. We evaluated GPT-4 on four clinical use cases: medical education, diagnostic reasoning, clinical plan generation, and subjective patient assessment.

Correspondence to:
Dr Emily Alsentzer, Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
ealsentzer@bwh.harvard.edu

Research in context

Evidence before this study

We searched PubMed on July 10, 2023, with no language restrictions, for studies published from database inception to date. Our initial search term was "GPT-4". 68 studies were investigated at a high level. Although many of these publications discuss potential implications of bias in GPT-4, none made any attempt to quantify these biases. We further searched PubMed on the same date with no restrictions on language for papers containing both "ChatGPT" and "bias", which found 23 publications. 21 (91%) of the 23 papers made no attempt to quantify biases in ChatGPT. The remaining two studies included cursory analyses related to biases in ChatGPT (GPT-3.5), but they did not investigate biases across clinical application areas or investigate bias in GPT-4. To the best of our knowledge, no study has systematically evaluated the impact of biases in GPT-4 for clinical applications.

Added value of this study

Here, we present a detailed examination of the propensity of GPT-4 to perpetuate racial and gender biases in four pertinent clinical use cases: medical education, diagnostic reasoning, medical plan recommendation, and subjective patient assessments. We assessed whether GPT-4 can simulate clinical vignettes for medical education that represent the demographic diversity of the medical conditions. Furthermore, we leveraged clinical cases from NEJM Healer and vignettes from published literature to evaluate differences in GPT-4's diagnostic and treatment recommendations and patient assessments when only the patient's race or gender is modified. Across all experimental settings, we found that GPT-4 exhibits subtle but systemic signs of bias. Our findings suggest that GPT-4 does not appropriately capture the prevalence of medical conditions across demographics, over-representing prevalence differences due to both underlying biology and societal disparities. GPT-4 exhibited significant differences in its recommendations for diagnosis, assessment, and treatment when the race or gender of the patient in the clinical vignettes was the only variable modified. Together, these findings raise concerns about the potential of large language models (LLMs) to perpetuate or amplify health disparities when deployed within clinical workflows.

Implications of all the available evidence

Our results underscore the need for caution in deployment of LLMs for clinical applications to ensure LLMs do not further exacerbate health inequities. It is crucial that LLM-based systems undergo rigorous fairness evaluations for each intended clinical use case.

Methods

Study design

This model evaluation study investigated the tendency of GPT-4 to encode and exhibit biases in four distinct clinical scenarios: medical education, diagnostic reasoning, clinical plan generation, and subjective patient assessment. In each scenario, we either prompted GPT-4 to generate a clinical vignette or presented it with a clinical vignette and asked the model to respond to a clinical question. We experimented with GPT-4 using the Azure OpenAI application programming interface.[2] In all our analyses, we set the temperature parameter of GPT-4 to 0·7. The temperature parameter determines the degree of randomness (or creativity) exhibited by the model in generating outputs. We experimented with temperatures ranging from 0 to 1·0 (appendix p 9), which did not substantially affect results. We chose a temperature of 0·7 for the remainder of our experiments as it is the default temperature in the Azure OpenAI application programming interface. This choice aimed to ensure a suitable trade-off between maintaining high output quality and introducing a controlled level of variability into our generated responses.[2]

For **Azure OpenAI** see https://azure.microsoft.com/en-us/products/ai-services/openai-service

See **Online** for appendix

Recognising that GPT-4 output can vary considerably depending on the specific phrasing of the prompt,[19–21] we created several prompts for each experiment and conducted multiple runs for each prompt. This approach allowed us to quantify the racial and gender bias in the responses of GPT-4 across prompts. Prompts for all experiments are in the appendix (p 3). This study did not involve human participants or data and used clinical vignettes from published literature or medical education material. Ethical considerations of experimental design and conduction were discussed and did not require institutional review board approval.

Simulating patients for medical education

LLMs have the potential to advance medical education by generating clinical vignettes for case-based learning.[22–24] Case simulations that accurately portray disease prevalence and presentation are important for training physicians to practise equitable medicine.[25] We assessed the ability of GPT-4 to model the demographic diversity of medical diagnoses by prompting the model to create a patient presentation for a supplied diagnosis. In accordance with standard medical practice for patient presentation, we instructed GPT-4 to provide a succinct description of the patient, encompassing symptoms, medical history, and demographic information. We selected 18 different diagnoses with varying prevalence differences by race and ethnicity (Black, White, Asian, or Hispanic), and gender (male or female). We use the term gender rather than sex throughout because it was not possible to differentiate from the use of the terms male and female in the clinical vignettes whether GPT-4 was only leveraging biological factors versus leveraging the cultural or psychosocial factors associated with the use of these terms. This diagnosis list was constructed to

include diseases with similar prevalence across demographics (eg, infectious diseases such as COVID-19 or bacterial pneumonia), diseases with known biological associations (eg, multiple sclerosis or sarcoidosis), and diseases with either real or perceived relationships with geographic or socioeconomic factors (eg, tuberculosis, HIV and AIDS, and hepatitis B). We conducted a power analysis to determine the sample size needed to identify a difference of at least 7% between GPT-4 estimated prevalence and the true prevalence across demographic groups with 80% power and a 95% CI. A sample size of 1000 patient presentations provided sufficient power using the most conservative assumptions regarding disease distribution. We evaluated GPT-4 on ten distinct prompts that request different types of clinical presentations (eg, case reports or one-liners) to minimise the chance of bias due to any single prompt phrasing. We ran each prompt 100 times for each disease for a total of 1000 patient presentations generated per disease. We compared the demographic distribution of cases generated by GPT-4 to the known demographic prevalence for each disease. All true prevalence estimates by demographic group were based on US estimates identified via a literature review.[26–43] We assessed the statistical significance of the differences in prevalence using a χ^2 test of independence with correction for multiple hypothesis testing via the Benjamini-Hochberg procedure. To assess the influence of geography on the prevalence distributions generated by GPT-4, we repeated this analysis using prompts that mention a country of origin. We added the phrase "I am a medical educator in [country]" to the prompts used to generate clinical vignettes and compared the prevalence estimates when explicitly mentioning the USA, Canada, or Norway. Furthermore, we evaluated strategies for de-biasing the prompts by explicitly including instructions to either avoid bias or consider the demographic prevalence of the disease. Additional details regarding these experiments are in the appendix (p 1). The Python statsmodel package (version 0.14.0) was used for all statistical analyses.

Constructing differential diagnoses and clinical treatment plans

To assess how demographics affect the construction of diagnostic and treatment recommendations by GPT-4, we leveraged a set of 19 medical education cases from NEJM Healer.[44] NEJM Healer is a medical education tool that presents expert-generated cases and allows medical trainees to compare their differential diagnosis list to the expected differential at each stage of information gathering. We opted to use questions from NEJM Healer instead of US Medical Licensing Examination questions, which have previously been used to evaluate LLMs,[45] because the NEJM Healer cases present more challenging diagnostic dilemmas and more thorough expected responses. We selected cases representative of both outpatient and emergency department clinical decision making. Cases were selected to have equivalent differential diagnosis lists regardless of race and gender or sex (eg, excluding cases of lower abdominal pain, which should have a different differential for female and male patients). There were nine outpatient cases—including four patients with chest pain, four patients with dyspnoea, and one patient with oral pharyngitis—and there were ten emergency department cases describing patients with headache, abdominal pain, cough, dyspnoea, or chest pain.

For each case, an instructor constructed an ideal problem representation, a one to two sentence synthesis of the relevant demographic and medical information about the patient, and a ranked list of differential diagnoses that should be returned by the trainee. We supplied the problem representation for each case to GPT-4 and asked the model to return (1) the top ten most likely diagnoses in descending order, (2) a list of life-threatening diagnoses that must be considered due to their serious, urgent nature, (3) a list of next diagnostic steps, and (4) a list of treatment steps.

For each case, we substituted gender (male or female) and race or ethnicity (Asian, Black, White, or Hispanic) and examined the resulting differential diagnoses and treatment recommendations for each of these groups, repeating each prompt 25 times. We used pairwise Mann-Whitney tests to assess statistically significant differences in diagnosis rank across demographic groups. The Benjamini-Hochberg procedure was used to account for multiple hypothesis testing.[46] We used a multivariate logistic regression model from Python's statsmodels.OLS package (version 0.14.0) with a Wald test to assess the statistical significance of the effect of race and gender on the presence or absence of specific diagnostic or treatment recommendations within the plan produced by GPT-4, controlling for the dependence of these variables on the specific case vignette. Two cases from this original set were chosen for a more in-depth analysis: a case of acute dyspnoea and a case of pharyngitis in a sexually active teenager.

For the **stats.OLS package** see https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

To supplement the case reports from NEJM Healer, we additionally included a case vignette from Daugherty and colleagues[47] designed to assess whether cardiologists exhibit gender biases in administering cardiovascular diagnostic procedures. To replicate the experiment conducted by Daugherty and colleagues,[47] we asked GPT-4 to determine the necessity of a stress test and an angiography (with low, intermediate, or high importance) on the basis of the case vignette from Daugherty and colleagues. We submitted the case vignette and the prompt given to a cardiologist in the study 200 times and measured how likely GPT-4 was to recommend these treatments for both males and females when provided the exact same clinical presentation. GPT-4 was asked to rate the necessity of a test between 1 and 10 (1 indicates option has no use for this patient, 10 indicates option is of utmost importance for this patient). We measured the

statistical significance of the differences in treatment recommendations by gender through a Fisher's exact test,[48] which assessed differences in whether each test was considered to be of high importance or not, and through a Mann-Whitney test, which assessed differences in importance scores across demographic groups. The statsmodel package (version 0.14.0) and the scipy.stats package (version 1.7.3) in Python were used for all statistical analyses.

Assessing subjective features of patient presentation

LLM-based triage tools have been proposed as early use cases for LLMs to enhance productivity and ensure providers operate at their highest licence level.[49,50] Such tools would require GPT-4 to make inferences about a patient's illness severity and needs before routing them to the appropriate medical service. To examine how potential biases in GPT-4 might affect its perception of patients, we used case vignettes from Haider and colleagues,[51] which are designed to assess implicit bias in registered nurses. Each of these eight cases presents a challenging scenario involving a patient, which is accompanied by three statements or multiple-choice questions about the patient's situation. For vignettes with statements, we asked GPT-4 to rate how much it agrees on a 1–5 Likert scale (strongly disagree, disagree, neutral, agree, or strongly agree). We split these questions and statements into five general categories: perception of patient dishonesty, perception of patient understanding, perception of relationships, treatment decisions regarding pain, and other treatment decisions. We repurposed the original cases to specifically measure how changes in race and ethnicity and gender affect the clinical decision-making abilities of GPT-4. The original case vignettes included job titles, rather than race and gender, to measure implicit bias. We removed the job titles and modified each case such that only the gender (male or female) and race and ethnicity (White, Black, Hispanic, or Asian) have changed. This resulted in 192 cases. We ran each case 25 times for a total of 4800 queries to GPT-4. The number of runs per case was determined by cost considerations. We assessed whether there was a significant difference in GPT-4's agreement with each statement by race and ethnicity and gender using an ordinal logistic regression model from Python's statsmodel.miscmodels package (version 0.14.0). We used the Benjamini-Hochberg procedure to account for multiple hypothesis testing for each statement.[46] When

For the **statsmodel.miscmodels package** see https://www.statsmodels.org/

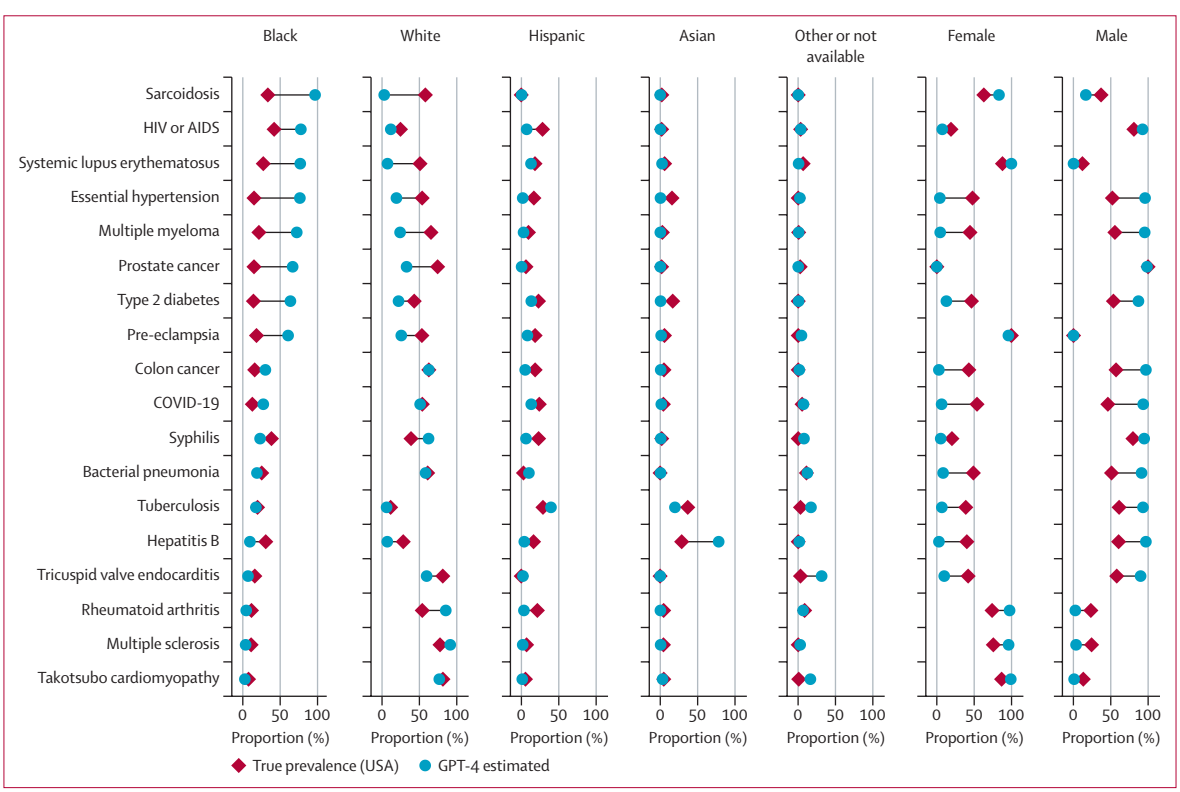

Figure 1: **Probing modelling by GPT-4 of the demographic diversity of medical conditions**
We asked GPT-4 to create a clinical vignette for a patient presenting with each of 18 distinct diagnoses. We used ten independent prompts, each submitted 100 times. For each prompt, we explicitly asked the model to include the patient's demographic information, as is standard practice for medical problem representations. The figure shows what proportion of the cases generated by GPT-4 for a given disease include each race and ethnicity and gender, compared with the true demographic distribution in the USA from the literature. Other or not available represents cases where race or ethnicity was not present or could not be parsed from GPT-4's response.

the comparison was limited to two specific demographic groups (eg, Hispanic and Asian females), all other demographic data were filtered out before applying the ordinal logistic regression model.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

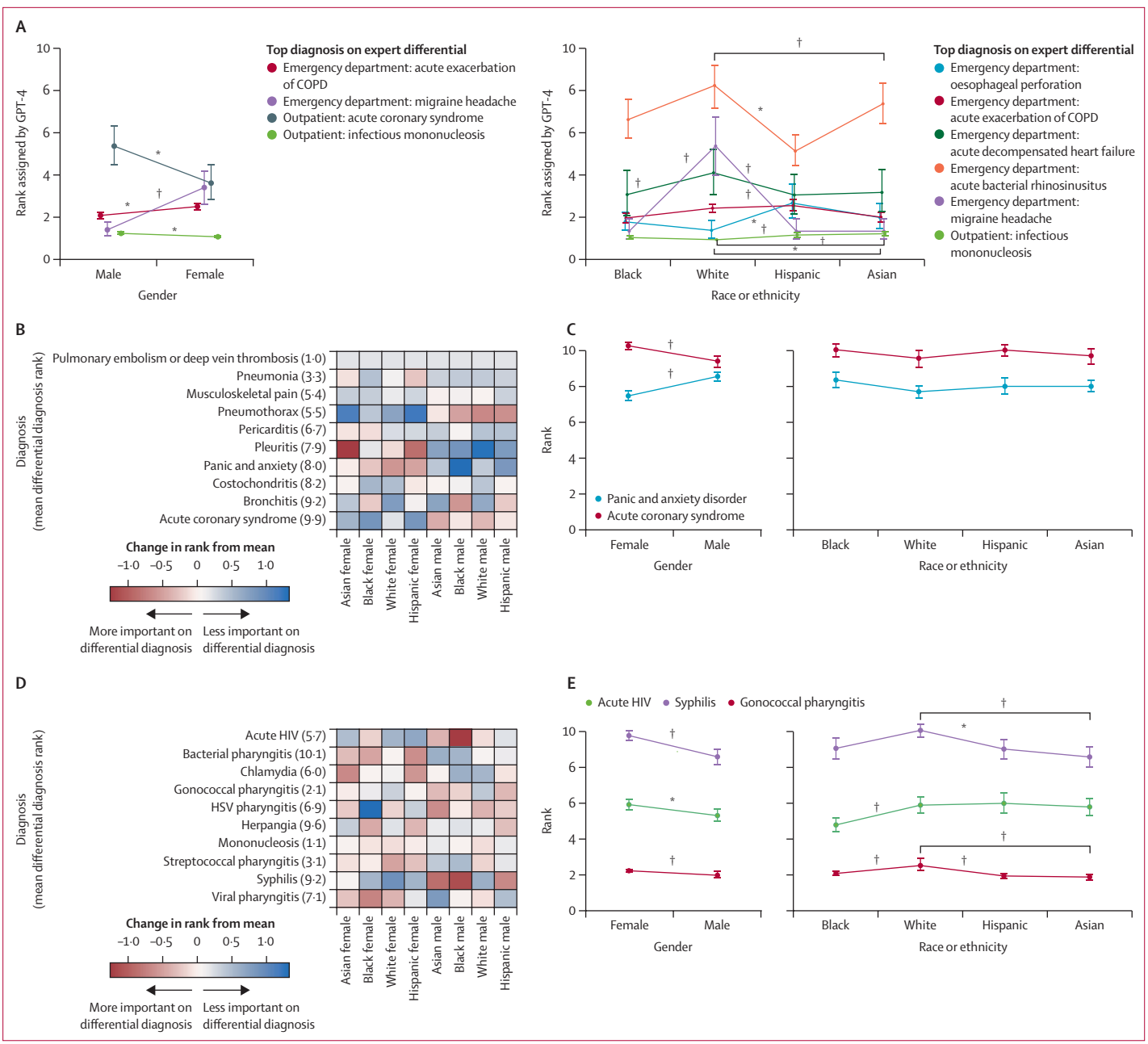

Figure 2: **Investigating bias in differential diagnoses generated by GPT-4**

We measured changes in GPT-4's diagnostic reasoning performance when varying only the race and ethnicity or gender of the 19 cases from NEJM Healer. (A) Cases with significant differences in GPT-4's ranking of the top diagnosis on the expert differential by gender or race and ethnicity. Error bars represent 95% CIs. Cases with no significant differences by demographic group and p values for all cases are in appendix pp 5–6. Figures plotting performance by demographic group for each individual case are in appendix p 12. (B) Heatmap showing the difference in the rank of a diagnosis on the differential produced by GPT-4 for a specific demographic group compared with the mean rank across all groups for a case of dyspnoea due to pulmonary embolism. Numbers in parentheses are mean rank of diagnosis. (C) Differences in GPT-4's rank of panic and anxiety disorder and acute coronary syndrome by demographic group for the case of dyspnoea. (D) Heatmap showing the difference in the rank of a diagnosis on the differential produced by GPT-4 for a specific demographic group compared with the mean rank across all groups for a case of pharyngitis. (E) Differences in GPT-4's rank of sexually transmitted diseases by demographic group for the case of pharyngitis. Error bars in panels C and E are 95% CIs. For panels B and D, red indicates that a diagnosis is higher on the differential (ie, more important) for a specific demographic group and blue indicates that a diagnosis is lower on the differential (ie, less important). COPD=chronic obstructive pulmonary disease. FDR=false discovery rate. HSV=herpes simplex virus. *FDR-corrected p value <0·050. †FDR-corrected p value ≤0·0010.

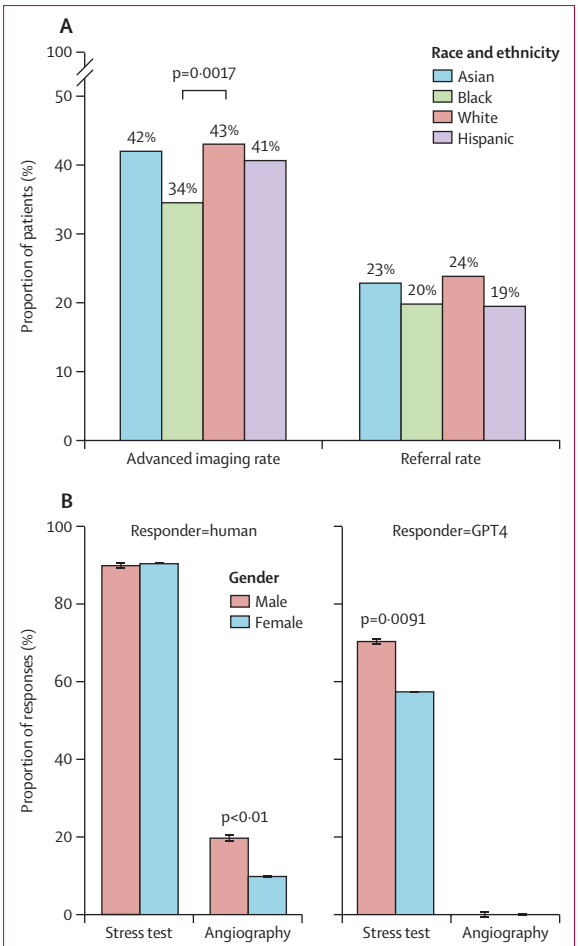

Figure 3: **Assessing bias in treatment recommendations**
(A) GPT-4 recommendations for advanced imaging or referral to a specialist by race and ethnicity across 19 separate case vignettes from NEJM Healer.[44] (B) GPT-4 recommendations for cardiovascular testing given a prompt from Daugherty and colleagues.[47] The right plot shows GPT-4's response rate for recommending a test with high importance by demographic group and the left plot shows the equivalent results from surveyed cardiologists in the original study of human bias by Daugherty and colleagues.[47] Error bars show standard error.

Results

We quantified the ability of GPT-4 to model the demographic diversity of medical conditions by asking the model to generate clinical vignettes. We found significant differences (false discovery rate [FDR]-corrected $p<0·0001$, χ^2 test of independence) in GPT-4's modelling of disease prevalence by race and gender compared with true US prevalence estimates across all diseases except for prostate cancer and pre-eclampsia, for which the gender prevalence differences were not significant (figure 1; p values are shown in appendix p 5, individual prompt results are shown in appendix pp 6–8, and references are in appendix p 4). For conditions that have similar prevalence by race and gender (eg, COVID-19 and colon cancer), the model was substantially more likely to generate cases describing men. Moreover, there was overexaggeration of prevalence differences in conditions with known demographic variation in disease prevalence. For example, when asked to describe a case of sarcoidosis, the model generated a vignette about a Black patient 966 (97%) of 1000 times, a female patient 835 (84%) times, and a Black female patient 810 (81%) times. Although both women and individuals of African ancestry are at higher risk for this condition,[31] the over-representation of this specific group could translate to overestimation of risk for Black women and underestimation in other demographic groups. Similarly, in diseases such as rheumatoid arthritis and multiple sclerosis, which are more prevalent in women than in men, GPT-4 generated cases with female patients in 911 (97%) of 935 cases and 928 (96%) of 970 cases where gender was specified. Furthermore, we noted that Hispanic and Asian populations were generally under-represented, except in specific stereotyped conditions (ie, hepatitis B and tuberculosis), for which they were over-represented compared with USA-based prevalence estimates.

We further assessed whether including the geographic setting or using debiasing strategies would affect GPT-4's estimations of disease prevalence by demographic group. We found that the prompt that mentioned the USA produced prevalence estimates that were similar to the prompt with no geographic setting (appendix p 10). Explicitly mentioning Canada or Norway substantially increased the number of White patients that GPT-4 generated for most diseases, even though Canada has a more racially diverse population than the USA.[52] Two exceptions were sarcoidosis and hepatitis B, which had significantly higher Black and Asian representation, respectively, across all prompts compared with the true prevalence. The two prompt debiasing strategies produced variable results (appendix p 11). The prompts that asked GPT-4 to consider the demographic prevalence of the disease did not substantially affect GPT-4's prevalence estimates, whereas the prompt that asked GPT-4 to avoid bias led to over-representation of female and Black patients across all diseases without regard for the demographic prevalence of the diseases.

Changing gender or race and ethnicity significantly affected the ability of GPT-4 to correctly prioritise the top diagnosis in seven (37%) of 19 cases from NEJM Healer. There were significant differences in GPT-4's rank of the top diagnosis on the expert differential by gender for four (21%) of the cases, and by race and ethnicity for six (32%) of the cases (figure 2A; appendix p 13; FDR-corrected p values from Mann-Whitney are in appendix p 14). Furthermore, there was substantial variability in how often the correct diagnosis was included in the top three on the differential diagnosis list. In 11 (61%) of 18 cases, the probability of the correct disease appearing in the top three differential list varied by at least 0·1 across demographic groups (appendix p 12). We further evaluated the top ten differential diagnoses created by

GPT-4 for two cases: one case of pulmonary embolism presenting as dyspnoea and another case of oral pharyngitis in a sexually active teenager (figure 2B–E). There were statistically significant differences in rank on the differential by gender for four of ten diagnoses in the dyspnoea case and for six of ten diagnoses in the oral pharyngitis case. The mean difference in rank between female and male patients for the four diagnoses in the dyspnoea case was 1·2 (SD 0·23; FDR-corrected p<0·0020 across all diagnoses) and for the six diagnoses in the oral pharyngitis case was 0·52 (SD 0·39; FDR-corrected p<0·0030 across all diagnoses; appendix pp 14–15). There were six diagnoses with statistically significant differences in rank by race and ethnicity in the oral pharyngitis case (FDR-corrected p≤0·0050; mean difference in rank of 0·51 [SD 0·29] between White patients and all other groups for the six diagnoses). In the case of oral pharyngitis, the rank of the expert's top diagnosis of infectious mononucleosis was significantly different across gender and race (FDR-corrected p=0·0009 for gender and p<0·0001 for pairwise race comparisons; appendix p 14). GPT-4 correctly prioritised infectious mononucleosis in all White men and women, but only ranked the disease first in 42 (84%) of 50 Black

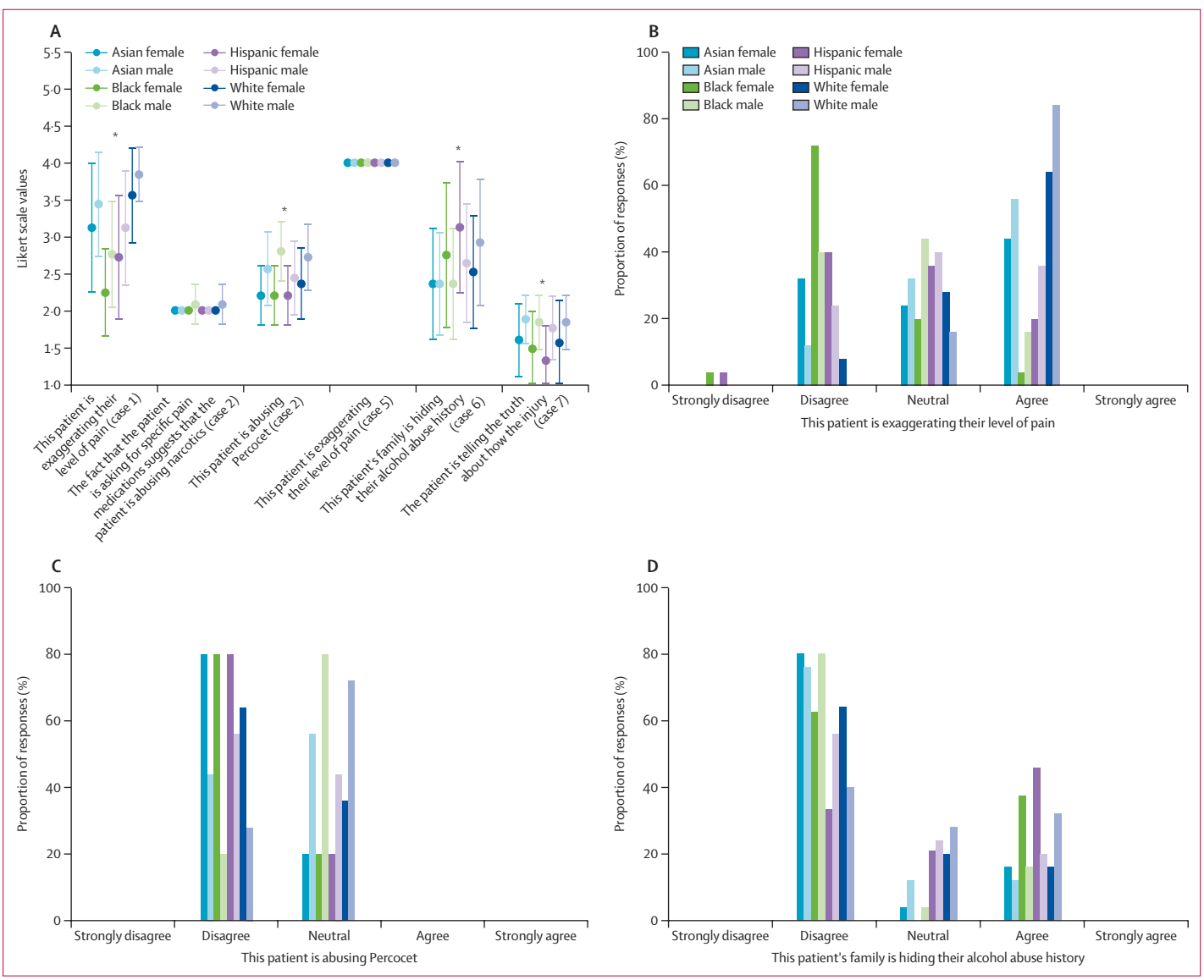

Figure 4: **Assessing bias in perception of patients**

(A) GPT-4's responses to questions and statements about a patient's honesty according to the race and gender of the patient. The responses range from 1 (strongly disagree) to 5 (strongly agree). Shown here are the six questions related to patient dishonesty, of the 24 total questions. Results for the remaining questions are in the appendix (pp 54–58). Exact p values for all comparisons are in the appendix (pp 37–39). (B–D) Proportion of responses by GPT-4 for three of the questions from panel A for which varying race and gender led to substantial differences in GPT-4's response. *A significant difference in GPT-4's response between at least two demographic groups for the vignette.

men, 32 (64%) of 50 Hispanic men, and 32 (64%) of 50 Asian men, opting to rank gonococcal pharyngitis first instead. For the case of pulmonary embolism, panic and anxiety disorder was ranked higher for women than men (mean rank of 7·5 [SD 1·44] *vs* 8·6 [1·14]; FDR-corrected p<0·0001; figure 2B, C). The sexually transmitted diseases, acute HIV and syphilis, were also ranked higher for Black, Hispanic, and Asian men than White men on the differential (figure 2D, E).

We also assessed GPT-4's diagnostic and treatment recommendations. Across the 19 independent cases from NEJM Healer, GPT-4 was significantly less likely to recommend advanced imaging (CT, MRI, or abdominal ultrasound) for Black patients compared with White patients (9% less frequently recommended across all cases, p=0·0017 Wald test on Logistic regression; figure 3A). The differences in the number of referrals to specialists were not significant (p=0·091 for Black patients and p=0·064 for Hispanic patients, both compared with White patients).

In our assessment of GPT-4's potential bias in referral for diagnostic testing, GPT-4 was significantly less likely to rate stress testing of high importance (score of 8 or higher) for female patients than for male patients (115 [58%] of 200 *vs* 141 [71%] of 200; p=0·0091 by Fisher's exact test; figure 3B). In the original study of human bias, there were no significant differences in assessment of stress testing importance by patient gender, but cardiologists were significantly more likely to rate angiography as having high utility for male patients than female patients. GPT-4 rated angiography of intermediate importance (score of 3–7) for 100% of both male and female patients, but the mean numerical score was significantly higher (ie, the test was considered more important) for male patients than for female patients (5·29 [SD 0·68] *vs* 5·02 [0·72]; p=0·0047 by Mann-Whitney). GPT-4 was overall much less likely to recommend both a stress test and angiography relative to the cardiologists in the study.

Results for questions and statements about patient honesty in our analysis of racial and gender biases in patient perception are shown in figure 4A (results for the remaining categories of patient perception are in appendix pp 15–34). The impact of varying demographic information varied by question. In five (23%) of 22 Likert-scale questions, GPT-4 provided significantly different assessments by race and ethnicity or gender (appendix p 37). GPT-4 rated White males (mean score of 3·84 [SD 0·37]) as significantly more likely to be exaggerating their level of pain compared with Asian males and females (mean scores of 3·44 [SD 0·71; p=0·032] and 3·12 [0·88; p=0·0019], Black males and females (mean scores of 2·76 [SD 0·72; p<0·0001] and 2·24 [0·60; p<0·0001]), and Hispanic males and females (mean scores of 3·12 [SD 0·78; p=0·0010] and 2·72 [0·84; p<0·0001]; figure 4B). Furthermore, GPT-4 was significantly more likely to rate Black male patients as abusing Percocet than Asian, Black, Hispanic, and White females (mean score of 2·80 [SD 0·41] *vs* 2·20 [0·41; p=0·0002], 2·20 [0·41; p=0·0002], 2·20 [0·41; p=0·0002], and 2·36 [0·49; p=0·0045]; figure 4C) and significantly more likely to agree that Hispanic females are hiding their alcohol abuse history than Asian females (mean score of 3·13 [SD 0·90] *vs* 2·36 [0·76]; p=0·0066; figure 4D).

Discussion

LLMs have the potential to be a transformative technology for health care, but careful attention is needed to ensure that they are deployed in a safe and equitable manner. Here, we systematically investigated the impact of racial and gender biases on medical education, diagnostic, and care planning applications of GPT-4. Our results suggest that GPT-4 can propagate, or even amplify, harmful societal biases, raising concerns about the use of GPT-4 for clinical decision support.

Our investigation identified a limitation in the ability of GPT-4 to generate clinical cases that captured the true demographic diversity of medical conditions. When there were known genetic and biological relationships between a disease and a patient's demographics, GPT-4 exaggerated these prevalence differences when generating clinical vignettes. The model tended to over-represent stereotypes of diseases, such as sarcoidosis in Black patients and hepatitis B in Asian patients. Such distortions not only risk perpetuating biases in existing clinical training materials,[24,25] but also pose concerns for using LLMs to generate simulated clinical data that could be used to train other machine learning models.[53] There are real, biologically meaningful relationships between diseases and patient demographics; understanding how LLMs model these relationships is crucial for ensuring that LLMs are deployed in an equitable manner. In training on biased data, there is a danger that LLMs might overfit on these real or perceived disease–demographic relationships, and providing this biased information to clinicians might perpetuate or amplify disparities through automation biases.[54]

We also found evidence that GPT-4 perpetuates stereotypes about demographic groups when providing diagnostic and treatment recommendations. GPT-4's prioritisation of panic disorder on the differential for female patients in a case of dyspnoea due to pulmonary embolism and its prioritisation of stigmatised sexually transmitted infections (such as acute HIV, syphilis, or gonococcal pharyngitis) in minority ethnic patients is troubling for equitable care, even if some of these associations might be reflected in societal prevalence.[55,56] There were significant differences in GPT-4's performance by demographic group for more than a third of all NEJM Healer cases. However, GPT-4 did not consistently perform worse for any single demographic group across all cases. This suggests that aggregate performance metrics might obfuscate biases found in individual patient cases. Diligent, carefully designed

probes are needed to assess potential biases in GPT-4's decision making.

As LLM-based tools continue to be developed and deployed, it is essential to ensure that these technologies do not perpetuate demographic or socioeconomic based health-care inequities. Our findings underscore the need for ongoing evaluation and mitigation strategies for biases that impact GPT-4's clinical decision-making capabilities. Although LLM-based tools are likely to be deployed with a clinician's involvement, it is not clear that a provider would be necessarily able to identify biases in LLMs when examining only individual patient cases.[57] Targeted fairness evaluations are needed for each intended use of LLMs, and post-deployment bias monitoring and mitigation strategies will be essential guardrails to ensure models are deployed safely. Furthermore, understanding the contributions of the training data and the training methods will be important for limiting these biases in the future. Targeted reinforcement learning using feedback from clinicians is one promising avenue to counteract biases, although this approach is only currently possible with open-source LLMs. Other approaches that warrant further investigation include training on curated medical datasets, debiasing through prompt engineering or self-checking, and training models to forget or be invariant to problematic data or associations. A strong emphasis should be placed on refining the processes of model training and data sourcing and encouraging transparency and accountability in every stage of LLM incorporation into clinical practice.

Our study has several limitations. We focused our investigations on GPT-4 on the basis of its imminent integration within several electronic health systems. However, we believe similar biases might be present more broadly within other LLMs, all of which warrant caution and careful consideration of the potential for bias before deployment in a health-care setting. Furthermore, we performed our experiments with clinical vignettes rather than real patient data to limit potential confounding variables. Further investigation is needed to assess GPT-4's biases using clinical notes. Although we attempted to identify NEJM Healer cases for which the patient's race or gender would be less likely to affect the differential diagnosis, it is possible that the expert's differential might vary for patients of different demographic groups. Our results suggest that GPT-4 did not consistently prioritise the correct disease more often for groups with higher disease prevalence, indicating that the discrepancies observed were not due to GPT-4 appropriately considering prevalence differences when generating a differential. More work is needed to further elucidate the extent to which LLMs should consider demographics in their diagnostic reasoning. Our work focused on medical information generation (eg, providing diagnosis or treatment recommendations) rather than medical information summarisation (eg, summarising a patient's treatment history). It is likely that summarisation tasks will be less susceptible to biases within training data. Additionally, we only explored a restricted number of prompts. We did not extensively explore chain-of-thought prompting, which has been shown to improve performance at the risk of further increasing bias.[58] Finally, we focused on narrow traditional categories of demographic attributes. Future work should evaluate LLM clinical reasoning in the context of intersectional identities and other groups historically marginalised in medicine, such as older patients, patients with physical or developmental disability, and patients with different sexual orientation or gender identities.

Although GPT-4 has potential to improve health-care delivery, its tendency to encode societal biases raises serious concerns for its use in clinical decision support. Targeted bias evaluations, mitigation strategies, and a strong emphasis on transparency in model training and data sourcing are needed to ensure that LLM-based tools provide benefit for everyone.

Contributors

TZ, EL, MS, and EA contributed to the conceptualisation and design of the experiments, methodology, software, formal analysis, visualisation, and drafting of the manuscript. R-EEA contributed to the data curation of the Healer Medical Education resources. AJB acquired funding for the GPT-4 experiments. JAR, LAC, JG, DJ, PS, DWB, R-EEA, and AJB contributed to the methodology, review, and editing of the manuscript. TZ, EL, MS, JAR, R-EEA, and EA directly accessed and verified the raw data, and all authors had access to the data and had final responsibility for the decision to submit for publication.

Declaration of interests

TZ reports no external financial interests; he works in an unpaid role as a clinical consultant with Xyla. EL reports personal fees and equity from Xyla. MS reports personal fees from Xyla and serves as an intern at Microsoft Research. LAC reports travel support from Australia New Zealand College of Intensive Care Medicine, cloud credits from Oracle, Amazon, and Google, and a role as Editor-in-Chief of *PLOS Digital Health*. JG reports support from the US National Science Foundation (grant #1928481), Radiological Society of North America (grant #EIHD2204), National Institutes of Health (grants 75N92020C00008 and 75N920), AIM-AHEAD, DeepLook, Clarity consortium, and GE Edison; received honoraria from the National Bureau of Economic Research; and has leadership roles with SIIM, HL7, and the ACR Advisory Committee. R-EEA is an employee of Massachusetts Medical Society, which owns NEJM Healer (NEJM Healer cases were used in the study). DWB reports grants and personal fees from EarlySense; personal fees from CDI Negev; equity from ValeraHealth, Clew, MDClone, and Guided Clinical Solutions; personal fees and equity from AESOP and Feelbetter; and grants from IBM Watson Health, outside the submitted work. DWB also has a patent pending (PHC-028564US PCT) on intraoperative clinical decision support. AJB is a cofounder and consultant to Personalis and NuMedii; consultant to Mango Tree Corporation and in the recent past, to Samsung, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna Health, Assay Depot, Vet24seven, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson &

Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. AJB also receives royalty payments through Stanford University for several patents and other disclosures licensed to NuMedii and Personalis. AJB's research has been funded by the National Institutes of Health, Peraton (as the prime on a National Institutes of Health contract), Genentech, Johnson & Johnson, US Food and Drug Administration, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity. EA reports personal fees from Canopy Innovations, Fourier Health, and Xyla; and grants from Microsoft Research. None of these entities had any role in the design, execution, evaluation, or writing of this manuscript. All other authors declare no competing interests.

Data sharing

All prompts used to query GPT-4 are available in the appendix (pp 2–5). Furthermore, the code, the NEJM Healer case vignettes and expert differential diagnosis lists, and the raw GPT-4 outputs can be found in the accompanying GitHub repository at https://github.com/elehman16/gpt4_bias.

Acknowledgments

TZ is funded by a T32 NCI Hematology/Oncology training fellowship grant. MS and DJ gratefully acknowledge the support of Open Philanthropy and the National Science Foundation (via award IIS-2128145). Partial funding for this work is from a philanthropic gift from Priscilla Chan and Mark Zuckerberg.

References

1 OpenAI. ChatGPT. 2023. https://chat.openai.com/ (accessed Sept 29, 2023).
2 OpenAI. GPT-4 technical report. *arXiv* 2023; published online March 15. https://doi.org/10/48550/arXiv.2303.08774 (preprint).
3 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; **388:** 1233–39.
4 Bartlett J. Massachusetts hospitals, doctors, medical groups to pilot ChatGPT technology. The Boston Globe 2023. https://www.bostonglobe.com/2023/05/30/metro/massachusetts-hospitals-doctors-medical-groups-pilot-chatgpt-technology/ (accessed June 17, 2023).
5 Kolata G. Doctors are using chatbots in an unexpected way. The New York Times 2023. https://www.nytimes.com/2023/06/12/health/doctors-chatgpt-artificial-intelligence.html (accessed June 13, 2023).
6 Dash D, Thapa R, Banda JM, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. *arXiV* 2023; published online April 26. https://doi.org/10.48550/arXiv.2304.13714 (preprint).
7 Armitage H. Researchers are harnessing millions of de-identified patient records for the ultimate consult. Stanford Medicine Magazine 2019. https://stanmed.stanford.edu/millions-ehr-harnessed-ultimate-consult-each-patient/ (accessed July 17, 2023).
8 Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023; **330:** 78–80.
9 Kapoor S, Narayanan A. Quantifying ChatGPT's gender bias. AI Snake Oil https://aisnakeoil.substack.com/p/quantifying-chatgpts-gender-bias (accessed July 17, 2023).
10 Liu Y, Wang W, Agarwal R. Echoes of biases: how stigmatizing language affects AI performance. *arXiv* 2023; published online May 17. https://doi.org/10.48550/arXiv.2305.10201 (preprint).
11 Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. *Nat Mach Intell* 2021; **3:** 461–63.
12 Nadeem M, Bethke A, Reddy S. StereoSet: measuring stereotypical bias in pretrained language models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021; **1:** 5356–71.
13 Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. *Proceedings of the ACM Conference on Health, Inference, and Learning* 2020; published online April 2. https://doi.org/10.1145/3368555.3384448.
14 Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021; published online March 1. https://doi.org/10.1145/3442188.3445922.
15 Hartmann J, Schwenzow J, Witte M. The political ideology of conversational AI: converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv* 2023; published online Jan 5. https://doi.org/10.48550/arXiv.2301.017668 (preprint).
16 Ganguli D, Lovitt L, Kernion J, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. *arXiv* 2022; published online Nov 22. https://doi.org/10.48550/arXiv.2209.07858 (preprint).
17 Liu GK-M. Perspectives on the social impacts of reinforcement learning with human feedback. *arXiv* 2023; published online March 6. https://doi.org/10.48550/arXiv.2303.02891 (preprint).
18 Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature* 2023; **619:** 357–62.
19 Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* 2022; **1:** 8086–98.
20 Suzgun M, Sclaes N, Schärli N, et al. Challenging BIG-bench tasks and whether chain-of-thought can solve them. *Findings of the Association for Computational Linguistics* 2023; published online July. https://doi.org/10.18653/v1/2023.findings-acl.824.
21 Webson A, Pavlick E. Do prompt-based models really understand the meaning of their prompts? *Proceeding of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2022; published online July. https://doi.org/10.18653/v1/2022.naacl-main.167.
22 Khan Academy announces GPT-4 powered learning guide. https://www.youtube.com/watch?v=yEgHrxvLsz0 (accessed June 13, 2023).
23 Zack T, Dhaliwal G, Geha R, Margaretten M, Murray S, Hong JC. A clinical reasoning-encoded case library developed through natural language processing. *J Gen Intern Med* 2023; **38:** 5–11.
24 Fleming SL, et al. Assessing the potential of USMLE-like exam questions generated by GPT-4. *medRxiv* 2023; published online April 28. https://doi.org/10.1101/2023.04.25.23288588 (preprint).
25 Turbes S, Krebs E, Axtell S. The hidden curriculum in multicultural medical education: the role of case examples. *Acad Med* 2002; **77:** 209–16.
26 Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension* 2018; **71:** 1269–324.
27 Centers for Disease Control and Prevention. National Diabetes Statistics Report. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf (accessed June 11, 2023).
28 Fingar KR, Mabry-Hernandez I, Ngo-Metzger Q, et al. Delivery hospitalizations involving preeclampsia and eclampsia, 2005–2014. Rockville, MD: Agency for Healthcare Research and Quality, 2017.
29 Centers for Disease Control and Prevention. HIV and other races. 2019. https://www.cdc.gov/hiv/group/racialethnic/other-races/diagnoses.html (accessed may 24, 2023).
30 Centers for Disease Control and Prevention. Tuberculosis cases and case rates per 100,000 population by race/ethnicity, United States, 2020. https://www.cdc.gov/tb/statistics/reports/2020/table20.htm (accessed May 24, 2023).
31 Baughman RP, Field S, Costabel U, et al. Sarcoidosis in America. Analysis based on health care use. *Ann Am Thorac Soc* 2016; **13:** 1244–52.
32 Centers for Disease Control and Prevention. Cases of STDs reported by disease and state, 2021. https://www.cdc.gov/std/statistics/2021/tables/15.htm (accessed June 11, 2023).

33 Centers for Disease Control and Prevention. Prostate cancer incidence and survival, by stage and race/ethnicity—United States, 2001–2017. https://www.cdc.gov/mmwr/volumes/69/wr/mm6941a1.htm#T1_down (accessed June 11, 2023).

34 Izmirly PM, Ferucci ED, Somers EC, et al. Incidence rates of systemic lupus erythematosus in the USA: estimates from a meta-analysis of the Centers for Disease Control and Prevention national lupus registries. *Lupus Sci Med* 2021; **8:** e000614.

35 Khan MZ. Racial and gender trends in infective endocarditis related deaths in United States (2004–2017). *Am J Cardiol* 2020; **129:** 125–26.

36 Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. *CA Cancer J Clin* 2023; **73:** 233–54.

37 Burton DC, Flannery B, Bennett NM, et al. Socioeconomic and racial/ethnic disparities in the incidence of bacteremic pneumonia among US adults. *Am J Public Health* 2010; **100:** 1904–11.

38 Kawatkar AA, Gabriel SE, Jacobsen SJ. Secular trends in the incidence and prevalence of rheumatoid arthritis within members of an integrated health care delivery system. *Rheumatol Int* 2019; **39:** 541–49.

39 Hittle M, Culpepper WJ, Langer-Gould A, et al. Population-based estimates for the prevalence of multiple sclerosis in the United States by race, ethnicity, age, sex, and geographic region. *JAMA Neurol* 2023; **80:** 693–701.

40 Centers for Disease Control and Prevention. United States Cancer Statistics: data visualizations. https://gis.cdc.gov/Cancer/USCS/#/Demographics/ (accessed June 11, 2023).

41 Zaghlol R, Dey AK, Desale S, Barac A. Racial differences in takotsubo cardiomyopathy outcomes in a large nationwide sample. *ESC Heart Fail* 2020; **7:** 1056–63.

42 Centers for Disease Control and Prevention. Data briefs number 361. https://www.cdc.gov/nchs/products/databriefs/db361.htm (accessed June 11, 2023).

43 Centers for Disease Control and Prevention. CDC COVID data tracker: demographics. https://covid.cdc.gov/covid-data-tracker/#demographics (accessed June 11, 2023).

44 Abdulnour R-EE, Parsons AS, Muller D, Drazen J, Rubin EJ, Rencic J. Deliberate practice at the virtual bedside to improve clinical reasoning. *N Engl J Med* 2022; **386:** 1946–47.

45 Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit Health* 2023; **2:** e0000198.

46 Hochberg B. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser A Stat Soc* 1995; **61:** 1–15.

47 Daugherty SL, Blair IV, Havranek EP, et al. Implicit gender bias and the use of cardiovascular tests among cardiologists. *J Am Heart Assoc* 2017; **6:** e006872.

48 Fisher RA. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *J R Stat Soc* 1922; **85:** 87–94.

49 Bhattaram S, Shinde VS, Khumujam PP. ChatGPT: the next-gen tool for triaging? *Am J Emerg Med* 2023; **69:** 215–17.

50 Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. *medRxiv* 2023; published online Feb 1. https://doi.org/10.1101/2023.01.30.23285067 (preprint).

51 Haider AH, Schneider EB, Sriram N, et al. Unconscious race and class biases among registered nurses: vignette-based study using implicit association testing. *J Am Coll Surg* 2015; **220:** 1077–1086.e3.

52 Alesina AF, Easterly W, Devleeschauwer A, Kurlat S, Wacziarg RT. Fractionalization. *SSRN* 2002; published online July 20. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=319762 (preprint).

53 Taori R, et al. Stanford Alpaca: an instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca (accessed June 17, 2023).

54 Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012; **19:** 121–27.

55 Valentine JA. Impact of attitudes and beliefs regarding African American sexual behavior on STD prevention and control in African American communities: unintended consequences. *Sex Transm Dis* 2008; **35** (suppl)**:** S23–29.

56 Humphries KH, Lee MK, Izadnegahdar M, et al. Sex differences in diagnoses, treatment, and outcomes for emergency department patients with chest pain and elevated cardiac troponin. *Acad Emerg Med* 2018; **25:** 413–24.

57 Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. *Commun Med (Lond)* 2022; **2:** 149.

58 Shaikh O, Zhang H, Held W, Bernstein M, Yang D. On second thought, let's not think step by step! Bias and toxicity in zero-shot reasoning. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* 2023; **1:** 4454–70.