Original Research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 9: 1–7

DOI: 10.1177/20552076231170499
journals.sagepub.com/home/dhj
SAGE

Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health

Michael V. Heinz[1,2], Sukanya Bhattacharya[1], Brianna Trudeau[1], Rachel Quist[1], Seo Ho Song[1], Camilla M. Lee[1] and Nicholas C. Jacobson[1,2,3,4]

Abstract

Background: With a rapidly expanding gap between the need for and availability of mental health care, artificial intelligence (AI) presents a promising, scalable solution to mental health assessment and treatment. Given the novelty and inscrutable nature of such systems, exploratory measures aimed at understanding domain knowledge and potential biases of such systems are necessary for ongoing translational development and future deployment in high-stakes healthcare settings.

Methods: We investigated the domain knowledge and demographic bias of a generative, AI model using contrived clinical vignettes with systematically varied demographic features. We used balanced accuracy (BAC) to quantify the model's performance. We used generalized linear mixed-effects models to quantify the relationship between demographic factors and model interpretation.

Findings: We found variable model performance across diagnoses; attention deficit hyperactivity disorder, posttraumatic stress disorder, alcohol use disorder, narcissistic personality disorder, binge eating disorder, and generalized anxiety disorder showed high BAC ($0.70 \leq BAC \leq 0.82$); bipolar disorder, bulimia nervosa, barbiturate use disorder, conduct disorder, somatic symptom disorder, benzodiazepine use disorder, LSD use disorder, histrionic personality disorder, and functional neurological symptom disorder showed low BAC ($BAC \leq 0.59$).

Interpretation: Our findings demonstrate initial promise in the domain knowledge of a large AI model, with performance variability perhaps due to the more salient hallmark symptoms, narrower differential diagnosis, and higher prevalence of some disorders. We found limited evidence of model demographic bias, although we do observe some gender and racial differences in model outcomes mirroring real-world differential prevalence estimates.

Keywords

Digital health, digital mental health, bias in mental health, artificial intelligence, digital mental health assessment

Submission date: 9 June 2022; Acceptance date: 31 March 2023

Introduction

Mental health disorders are highly prevalent and burdensome globally, with a considerable increase in prevalence and burden over the last three decades. Nearly 970 million people (1 in 8) suffer from a mental disorder worldwide,[1] and mental disorders account for 125 million disability-adjusted life years (DALYs); between 1990 and 2019, both prevalence and DALYs increased by nearly

[1]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA
[2]Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
[3]Quantitative Biomedical Sciences Program, Dartmouth College, Hanover, NH, USA
[4]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA

Corresponding author:
Michael V. Heinz, Center for Technology and Behavioral Health, Dartmouth College, 46 Centerra Pkwy, Lebanon, NH 03766, USA.
Email: michael.v.heinz@dartmouth.edu

50%, with mental disorders now accounting for 5% of total global DALYs.[1] Further, mental disorders impose a considerable economic burden, with an estimated global cost of US$ 2.5 trillion.[2]

The need for mental health care far outweighs available resources, with the global mental health treatment gap estimated at nearly 60%.[3] In addition, misdiagnosis rates range from 66% to 98% for common psychiatric disorders, including depression and anxiety, in primary care, which is often the first point of contact for patients with mental health complaints.[4] Given the prevalence and burden coupled with treatment gaps, there is clearly a need for high-quality, cost-effective, and scalable mental health assessments. Novel automated systems leveraging artificial intelligence (AI) may be a promising solution to these challenges.

Broadly defined, AI comprises computational systems that mimic human cognitive processes, such as learning, reasoning, problem-solving, pattern recognition, generalization, and predictive inference.[5] With technological advances driving an increase in both data storage and computational power, AI has the potential to bolster existing mental healthcare modalities,[6] and AI has demonstrated the capacity to aid in mental health assessment[5]; AI has the potential to identify mental illness at earlier stages when interventions can be more successful[7] and provide a means for more convenient, affordable care.[6]

Although AI systems have great potential for improving mental health assessment and intervention, little is known of their accuracy, reliability, and generalizability, necessitating considerable caution and exploration before implementation. Some researchers argue that large language models may even be dangerous, producing human-like, comprehensible, fluent language, without basis in common sense or rational judgment—"stochastic parrots,"[8] or "a mouth without a brain."[9] Within the mental health domain, there are concerns regarding the capacity of AI to detect nuanced clinical presentations and subtle forms of human interaction.[6] Further, such models, trained on implicitly biased real-world data, could perpetuate or amplify existing human biases,[10] of particular concern in mental health care. Although some diagnostic error is likely unavoidable, it is important to understand this error and how it compares to expert, "gold-standard" assessment.

Evidence suggests that the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) diagnostic reliability is variable with poor reliability for some common psychiatric diagnoses, including generalized anxiety disorder (GAD) and major depressive disorder (MDD).[11] Further, there is known demographic bias in expert mental health diagnoses, based on sexual orientation, gender identity, socioeconomic status, race, ethnicity, and other demographic features.[12] For instance, there is evidence of potential gender biases in (a) autism spectrum disorder, where men are diagnosed around four times more frequently than women[13] and (b) borderline personality disorder (BPD), where women are diagnosed three times more frequently than men.[14]

Given human diagnostic error and known biases, AI models trained with historical data would risk perpetuating existing problems, which may prove detrimental to patient health.[15] Thus, in order to avoid perpetuating existing bias, the potential for bias in AI needs to be explored prior to more widespread use.[16] Ultimately, if AI can be strategically trained and applied, it may even present an opportunity to mitigate bias with predictive diagnoses.[17]

The primary aim of the present work was to evaluate the mental health domain knowledge of a large, general AI model; to accomplish this aim, we evaluated the model's capacity to generate accurate clinical interpretations, given contrived clinical vignettes. The secondary aim was to explore the potential for demographic bias in the model. To accomplish this aim, we presented the model with vignettes having demographically varied subjects. We hypothesized (i) the model's performance would parallel real-world diagnostic reliability and (ii) the model would recapitulate real-world differential prevalence estimates among mental health disorders.

Methods

Composition of vignettes

We drafted clinical vignettes based on stylistic guidelines established by the National Board of Medical Examiners for test questions appropriate for the USMLE STEP 2 and Subject Examinations (specifically, psychiatry).[18] Vignettes were cross-validated with the latest diagnostic guidelines set forth by the DSM-5.[19] Thus, each vignette was drafted in a manner that permitted flexibility in diagnosis, assessment, and planning but with a single best answer. We wrote and used a total of 59 distinct clinical vignettes (count in parentheticals) comprising personality disorders (7), mood disorders (9), anxiety disorders (6), trauma-related disorders (4), substance use disorders (12), eating disorders (7), sleep disorders (4), psychotic disorders (3), attention and impulse control disorders (4), and somatic symptom disorders (3).

Programmatically varying vignette demographics

Using Python, we programmatically created all combinations of age groups (child, adolescent, young adult, middle-aged adult, and older adult), sex (female and male), and race (Asian American, Black, Latino, Native American, and White) for each clinical vignette. We manually excluded generated vignettes with illogical phase-of-life-age combinations (e.g. "Parents pick up their 87-year-old son from school"). We changed pronouns

programmatically to match the sex stated at the beginning of the vignette.

The generative pre-trained transformer 3

We applied for, and were granted, an academic license for the OpenAI generative pre-trained transformer 3 (GPT-3).[20] GPT-3 is a large (175 billion parameter), pre-trained predictive language model, capable of text-prompt autocompletion. For instance, GPT-3 might return "Anxiety" when prompted with "Name a common psychiatric disorder." Using the GPT-3 Application Programming Interface for Python,[21] we programmatically prompted the model with the clinical vignettes and recorded the model's text responses. We used prompt design guidelines[21] for GPT-3 autocompletion functionality, providing two example vignettes with diagnostic interpretation and four vignettes without diagnostic interpretation (see Figure 1). We used the following modeling settings: (a) temperature: 0.5–0.55, (2) max tokens: 300, (3) sampling: top P. By pilot testing, we found that these settings generally resulted in the best balance between returning consistent, semi-structured responses, while maintaining creativity and variability in responses. The model provided $N = 1710$ semi-structured responses, which were parsed both programmatically (using Python regular expressions) and manually, where necessary.

Statistical analysis and demographic models

Balanced accuracy for model diagnostic performance. Using both manual annotation and text-string match, we standardized each model interpretation to ensure we included abbreviations and variable representations of psychiatric disorders (e.g., attention deficit hyperactivity disorder (ADHD)). We calculated balanced accuracy (BAC) to measure performance, using the R Caret package.[22] Considering the variable reliability of mental health diagnoses,[11] we defined the following BAC thresholds: 0.50 to 0.59 was poor; 0.60 to 0.69 was fair; 0.70 to 0.79 was good; 0.80 and above was excellent.

Linear mixed-effects modeling. Using the linear mixed-effects models using "Eigen" and S4 (LME4) library for R,[23] we fit generalized linear mixed-effects models to demographic features (i.e. race, age, and sex) [fixed effects], processing features (i.e. identifier for a base vignette, identifier for GPT-3 input batch) [random effects], and presence of a diagnosis [outcome binary variable]. Age was scaled and centered; "White" was used as the reference value for race, and "Female" was used as the reference value for sex. Using the output from these models, we calculated odds ratios (ORs), 95% confidence intervals, and associated p-values for demographic variables. We considered $p < 0.05$ to be statistically significant.

Results

Model performance across psychiatric diagnoses is shown by BAC, displayed in Table 1. Additional performance metrics, including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are displayed to further characterize the model performance. We show a 95% confidence interval for each metric. Diagnostic BAC varied considerably [0.50, 0.82]. The following disorders showed excellent BAC: ADHD (BAC = 0.82) and posttraumatic stress disorder (PTSD) (BAC = 0.81); the following disorders showed good BAC: alcohol use disorder (AUD) (BAC = 0.75), narcissistic personality disorder (BAC = 0.72), binge eating disorder (BAC = 0.72), and GAD (BAC = 0.70); the following disorders had fair BAC: antisocial personality disorder (BAC = 0.69), MDD (BAC = 0.69), social anxiety disorder (BAC = 0.67), psychotic disorder (BAC = 0.66), stimulant use disorder (BAC = 0.65), BPD (BAC = 0.63), and anorexia nervosa (BAC = 0.61); the following disorders showed poor BAC: bipolar disorder (BAC = 0.58), bulimia nervosa (BAC = 0.57), barbiturate use disorder (BAC = 0.55), conduct disorder (BAC = 0.55), somatic symptom disorder (BAC = 0.54), benzodiazepine use disorder (BAC = 0.53), LSD use disorder (BAC = 0.53), histrionic personality disorder (BAC = 0.52), and functional neurological disorder (BAC = 0.50).

Supplemental Tables 1 to 7 show output from linear mixed-effects models, with demographic features as fixed effects, base vignette and batch as random effects, and diagnosis as the outcome. We present ORs, 95% confidence intervals, and associated p-values for mood disorders ($p < 0.05$ considered significant), anxiety disorders, substance use disorders, personality disorders, eating disorders, psychotic disorders, and PTSD.

We found that Latino persons are more likely to be diagnosed with any mood disorder compared to White persons (OR = 0.64 [0.41, 0.996], $p < 0.05$); Native American persons are more likely to be diagnosed with substance use disorders (OR = 1.94 [1.02, 3.66], $p < 0.05$) compared to White persons; Native American persons are more likely to be diagnosed with AUD (OR = 2.79 [1.12, 6.97], $p < 0.05$) compared to White persons; men are less likely to be diagnosed with BPD (OR 0.43 [0.22, 0.86], $p < 0.05$) compared to women.

Discussion

Given the high prevalence and major social and economic impacts of untreated mental health disorders coupled with the relative shortage of mental health providers, there is a need for exploration of scalable technologies to address mental health care. While AI provides a promising direction for addressing both mental health assessment and treatment,[5] large deep learning models are plagued by a lack

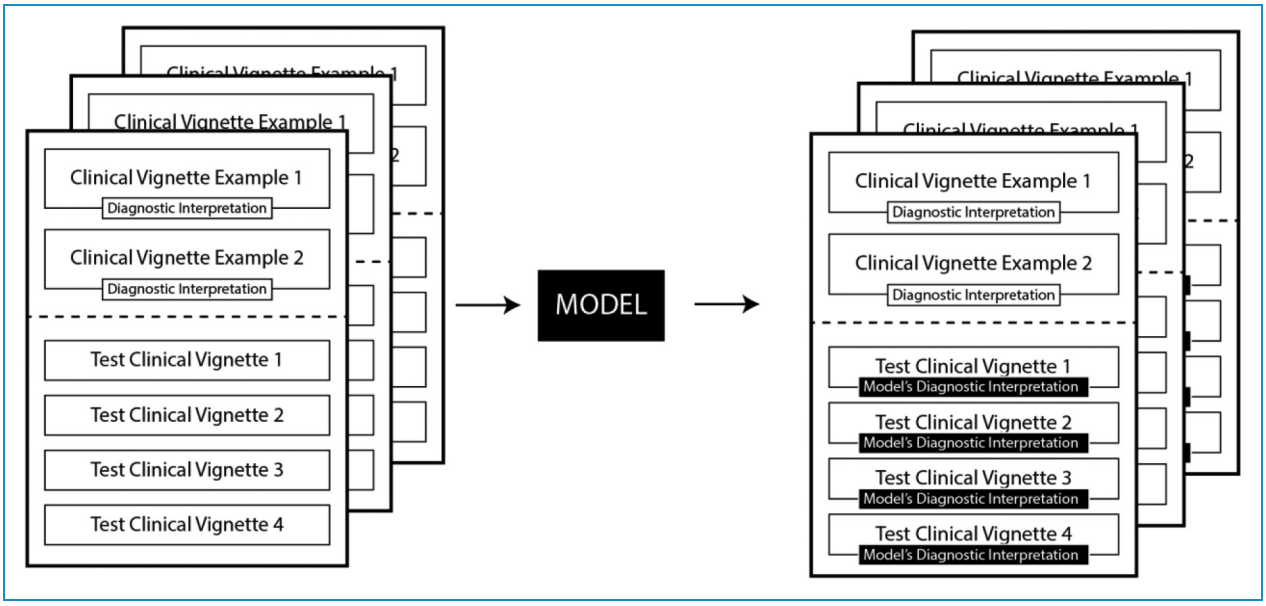

Figure 1. Schematic representation of prompt structure shown to the generative pre-trained transformer 3 (GPT-3) model. Each prompt included two clinical vignettes labeled diagnostic interpretations (i.e. a single best diagnostic label) and four clinical vignettes without a label. The model provided diagnostic interpretations for the four clinical vignettes without labeled diagnostic interpretations.

of transparency and interpretability,[24] potentially consigned to the same biases and inaccuracies of the data used to train them. Given their "black-box" nature, it is necessary to interrogate such models thoroughly to better understand generalizability, performance, and potential for bias prior to broad deployment.[16] In this study, we measured the performance and potential for bias of a large, generative AI model.

The generative model showed variable interpretive performance across clinical vignettes reflecting a broad range of mental health pathologies. The BAC for model diagnosis ranged from poor (BAC $\leq$ 0.59) to excellent (BAC $\geq$ 0.80). For all mental health problems, the model demonstrated superior specificity (Spe) and NPV compared to sensitivity (Sen) and PPV (see Table 1), suggesting the model would perform well at ruling out mental health disorders.

The model showed the best performance for ADHD, PTSD, and AUD. We hypothesized the cause for this to be readily identifiable salient keywords or concepts in the clinical vignettes for these disorders, which may also appear in the disorder name. For example, ADHD vignettes often included words such as "attention" and "distraction"; AUD included the word "alcohol" or the name of a specific alcoholic beverage; PTSD vignettes often included references to "flashbacks" or "nightmares." In contrast, the model performed the most poorly on histrionic personality disorder, functional neurological disorder (also known as conversion disorder), benzodiazepine use disorder, and LSD use disorder (BAC $\leq$ 0.53). We hypothesized this to be due in part to a broad, complex differential (in the case of functional neurological disorder and histrionic personality disorder), potentially including nonpsychiatric medical disorders (in the case of functional neurological disorder). Although LSD and benzodiazepine use disorder vignettes contain salient keywords (which we posit to have positively impacted the model's performance in AUD) we attribute the lower performance here to the considerably lower prevalence of these substance use disorders compared to AUD (19). The lower prevalence of these disorders may have translated to a relatively lower representation of such topics in the model training data, and thus the relatively lower likelihood of the model to detect these diagnoses.

Our findings may be partially contextualized in the DSM-5 Field Trials Part II,[11] examining the reliability of DSM-5 mental health diagnoses; the field trials found a wide distribution of reliability across psychiatric diagnoses. In particular, some common diagnoses with salient keywords or discrete prerequisite events (e.g., ADHD, PTSD, and AUD) showed relatively high test–retest reliability, characterized as "good" or "very good."[11] In contrast, some common disorders with wide differentials (e.g. MDD and GAD) had lower reliability, characterized as "questionable."[11] Although the field trial test–retest reliability scores are not directly comparable to the BAC metric used in our study, the field trials are useful in establishing a standard point of reference for model performance and perhaps tempering expectations for AI diagnosis detection.

Our secondary aim was to explore the potential for bias in the language model, considering this an early prerequisite to further translational development and deployment. In epidemiologic studies, prevalence estimates for mental health disorders vary between demographic groups, perhaps due to true prevalence differences, diagnostic

Table 1. Model performance across psychiatric diagnoses.

Disorder	BAC [0.95 CI]	Sen [0.95 CI]	Spe [0.95 CI]	PPV [0.95 CI]	NPV [0.95 CI]
ADHD	0.82 [0.75, 0.87]	0.64 [0.52, 0.75]	0.99 [0.99, 1.00]	0.79 [0.66, 0.89]	0.98 [0.98, 0.99]
Posttraumatic stress disorder	0.81 [0.75, 0.85]	0.64 [0.54, 0.74]	0.97 [0.96, 0.98]	0.59 [0.49, 0.69]	0.98 [0.97, 0.98]
Alcohol use disorder	0.75 [0.71, 0.79]	0.52 [0.44, 0.60]	0.98 [0.97, 0.98]	0.70 [0.60, 0.78]	0.95 [0.94, 0.96]
Binge eating disorder	0.72 [0.65, 0.78]	0.47 [0.34, 0.60]	0.98 [0.97, 0.98]	0.44 [0.31, 0.57]	0.98 [0.97, 0.99]
Narcissistic personality disorder	0.72 [0.61, 0.80]	0.43 [0.25, 0.63]	1.00 [0.99, 1.00]	0.76 [0.50, 0.93]	0.99 [0.98, 0.99]
Generalized anxiety disorder	0.70 [0.63, 0.76]	0.42 [0.29, 0.55]	0.97 [0.97, 0.98]	0.37 [0.26, 0.50]	0.98 [0.97, 0.99]
Major depressive disorder	0.69 [0.65, 0.73]	0.49 [0.42, 0.57]	0.89 [0.87, 0.90]	0.34 [0.29, 0.40]	0.94 [0.92, 0.95]
Antisocial personality disorder	0.69 [0.62, 0.75]	0.38 [0.26, 0.52]	1.00 [0.99, 1.00]	0.79 [0.60, 0.92]	0.98 [0.97, 0.98]
Social anxiety disorder	0.67 [0.63, 0.71]	0.35 [0.27, 0.44]	0.99 [0.99, 1.00]	0.79 [0.66, 0.89]	0.95 [0.94, 0.96]
Psychotic disorder	0.66 [0.61, 0.71]	0.33 [0.24, 0.44]	0.99 [0.98, 0.99]	0.60 [0.45, 0.74]	0.96 [0.95, 0.97]
Stimulant use disorder	0.65 [0.60, 0.70]	0.30 [0.21, 0.41]	1.00 [1.00, 1.00]	0.93 [0.77, 0.99]	0.96 [0.95, 0.97]
Borderline personality disorder	0.63 [0.58, 0.68]	0.28 [0.19, 0.38]	0.98 [0.98, 0.99]	0.50 [0.36, 0.64]	0.96 [0.95, 0.97]
Anorexia nervosa	0.61 [0.56, 0.67]	0.23 [0.13, 0.36]	0.99 [0.99, 1.00]	0.54 [0.33, 0.73]	0.97 [0.96, 0.98]
Bipolar disorder	0.58 [0.53, 0.62]	0.19 [0.11, 0.29]	0.97 [0.96, 0.97]	0.23 [0.14, 0.35]	0.96 [0.94, 0.96]
Bulimia nervosa	0.57 [0.53, 0.60]	0.13 [0.07, 0.22]	1.00 [0.99, 1.00]	0.80 [0.52, 0.96]	0.95 [0.94, 0.96]
Barbiturate use disorder	0.55 [0.49, 0.61]	0.10 [0.02, 0.27]	1.00 [0.99, 1.00]	0.50 [0.12, 0.88]	0.98 [0.98, 0.99]
Conduct disorder	0.55 [0.44, 0.65]	0.10 [0.00, 0.45]	1.00 [0.99, 1.00]	0.14 [0.00, 0.58]	0.99 [0.99, 1.00]
Somatic symptom disorder	0.54 [0.50, 0.58]	0.08 [0.03, 0.18]	0.99 [0.99, 1.00]	0.29 [0.10, 0.56]	0.97 [0.96, 0.97]
Benzodiazepine use disorder	0.53 [0.48, 0.59]	0.07 [0.01, 0.22]	1.00 [1.00, 1.00]	1.00 [0.16, 1.00]	0.98 [0.98, 0.99]
LSD use disorder	0.53 [0.48, 0.59]	0.07 [0.01, 0.22]	1.00 [1.00, 1.00]	1.00 [0.16, 1.00]	0.98 [0.98, 0.99]
Histrionic personality disorder	0.52 [0.47, 0.56]	0.03 [0.00, 0.17]	1.00 [0.99, 1.00]	0.25 [0.01, 0.81]	0.98 [0.98, 0.99]
FND[a]	0.50 [0.47, 0.53]	0.00 [0.00, 0.12]	1.00 [0.99, 1.00]	0.00 [0.00, 0.46]	0.98 [0.97, 0.99]

BAC: balanced accuracy; Sen: sensitivity; Spe: specificity; PPV: positive predictive value; NPV: negative predictive value; 0.95 CI: 0.95 confidence intervals are shown for all estimates; ADHD: attention deficit hyperactivity disorder.
[a]Functional neurological symptom disorder; sometimes interchanged with conversion disorder.

bias, demographically distinct disorder phenotypes, or differential levels of perceived stigma or accessibility.[12,13,25] Thus, we hypothesized that the AI model would emulate real-world differential prevalence estimates, given training on historical data.

The model displayed statistically significant biases for the following psychiatric disorders: BPD (male OR = 0.43 [0.22, 0.86], $p < 0.05$, reference = female), any mood disorder (Latino OR = 0.64, [0.41, 0.996], $p < 0.05$, reference = White), any substance use disorder (Native American OR = 1.94, [1.02, 3.66], $p < 0.05$, reference = White), and AUD (Native American OR = 2.79, [1.12, 6.97], $p < 0.05$, reference = White). Notably, the model replicated an established gender bias observed in BPD, with men being less likely to be diagnosed with BPD than women.[14] Further, the model reflected real-world

substance use disorder disparities, specifically the higher prevalence of illicit substance use disorders and AUD among American Indians and Alaska Natives.[26]

Despite evidence for real-world demographic biases in psychotic disorders (diagnosed more often in Black patients)[25] and ADHD (diagnosed more often in boys),[27] we did not find additional statistically significant differences in model prediction based on demographic variables (though trends are evident, see Supplemental Tables 1 to 7). This finding was inconsistent with our original hypothesis and should be contextualized by highlighting the following points: (a) the contrived vignettes were *identical* across demographic groups, except for single word substitutions to vary age, race, and sex; thus, we would expect very low variability across distinct demographic groups (b) the temperature (T) hyperparameter was limited ($0.5 < T < 0.55$) to produce consistent, semi-structured output, potentially at the cost of greater model creativity and variability; limiting this hyperparameter may have led to more consistent responses, potentially underestimating bias. (c) Though not statistically significant, our findings do reveal trends in the direction expected, based on real-world prevalence differentials (e.g. OR < 1 for MDD in men, OR < 1 for bipolar disorder in non-White persons, see Supplemental Table 1).

Given known bias in language models,[16] OpenAI has recently explored strategies to reduce systematic model bias, positing the process for adapting language models to society (PALMS).[28] PALMS leverages small, curated, values-based datasets to iteratively target observed shortcomings in topic categories involving injustice and inequality, mental health, U.S. protected groups, and others (see Appendix A).[28] The adjusted model showed significantly improved behavior, measured by toxicity scores assigned by human annotators.[28] These findings demonstrate an effective approach for addressing undesirable model behaviors, highlighting the utility of values-targeted curated data sets as an effective solution for mitigating toxicity and harmful bias in AI.

Strengths and limitations

Our research was the first to quantitatively explore the mental health domain knowledge and the potential for demographic bias of a large, general language model. We used a large, carefully curated dataset, representing broad mental health pathology and structured with the most up-to-date diagnostic criteria for mental disorders. Despite the aforementioned strengths, this study had several important limitations. (a) The clinical vignettes used to prompt the model were contrived, and though validated with DSM-5 diagnostic criteria, did not represent naturalistic clinical presentations. The vignettes were streamlined to present core diagnostic features and did not have all the subtleties and extraneous details expected in a real-world encounter. Despite these limitations, we believe this method allowed for (i) effective probing of the model's mental health domain knowledge in a practical context and (ii) exploring bias in a controlled context with the ability to hold all other potentially confounding variables constant. (b) We used a limited subset of the model's hyperparameters, notably temperature, which varied the model's creativity and randomness. Though we pilot-tested various hyperparameters, selecting the best performers, future research should compare performance and bias with systematically varied model hyperparameters. (c) We explored a relatively small subset of possible demographic variables and combinations, limited mainly by practical considerations (i.e. computational run time and cost). Future research should include additional racial-ethnic and gender categories (e.g. non-binary gender). (d) Due to similar practical considerations, we prompted the model with batches ($n = 4$) of vignettes, rather than single vignettes. This approach had the potential for the model conflating details from two distinct vignettes. To account for this, we modeled the batch identifier as a random effect.

Conclusion

AI-powered systems present a promising, scalable solution to the rapidly growing gap between the need and availability of mental health care. Even so, the novelty and often inscrutable nature of such systems, raise concern for their use in high-stakes domains, such as healthcare.[24] Exploratory measures to ensure the efficacy and equity of such systems are a necessary prerequisite for future deployment. Our work is the first to systematically and quantitatively address this challenge through the evaluation of the psychological knowledge base and exploration of potential bias in a large, general AI model. Though we find high variability in performance across diagnoses, our work suggests initial promise in domain knowledge with mild demographic bias; future work is needed to further assess the effectiveness of such models in naturalistic settings.

Acknowledgements: None.

Contributorship: MH and NJ conceived and designed the study and supervised the project. MH, RQ, CLSB, BT, and SS designed and validated mental health prompts; MH, RQ, CL, SB, BT, and SS parsed and structured model output in spreadsheet format. MH and NJ designed and executed statistical analysis. All authors contributed to manuscript drafting. MH, SB, BT, and RQ provided a literature search and critical manuscript review. MH and NJ approved the manuscript for publication.

Consent statement: The authors affirm that the study described in this manuscript did not involve human subjects, and as such, consent was not obtained.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The authors did not seek institutional ethics approval because the present study did not involve human subjects. The study was limited to the investigation of a pre-trained language model.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially funded by the National Institute of Mental Health (NIMH) and the National Institute Of General Medical Sciences (NIGMS) under 1 R01 MH123482-01.

Guarantor: MH.

ORCID iDs: Michael V. Heinz https://orcid.org/0000-0003-0866-0508
Seo Ho Song https://orcid.org/0000-0003-2970-2746

Supplemental material: Supplemental material for this article is available online.

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *Lancet Psychiatry* 2022; 9: 137–150.
2. Trautmann S, Rehm J and Wittchen H. The economic costs of mental disorders. *EMBO Rep* 2016; 17: 1245–1249.
3. Wang PS, Lane M, Olfson M, et al. Twelve-month use of mental health services in the United States: Results from the national comorbidity survey replication. *Arch Gen Psychiatry* 2005; 62: 629.
4. Vermani M, Marcus M and Katzman MA. Rates of detection of mood and anxiety disorders in primary care: A descriptive, cross-sectional study. *Prim Care Companion CNS Disord* 2011; 13: PCC.10m01013.
5. D'Alfonso S. AI in mental health. *Curr Opin Psychol* 2020; 36: 112–117.
6. Graham S, Depp C, Lee EE, et al. Artificial intelligence for mental health and mental illnesses: An overview. *Curr Psychiatry Rep* 2019; 21: 116.
7. Nemesure MD, Heinz MV, Huang R, et al. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci Rep* 2021; 11: 1980.
8. Bender EM, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency [Internet]. New York, NY, USA: Association for Computing Machinery; 2021 [cited 2022 Apr 3]. pp. 610–623.
9. Hutson M. Robo-writers: The rise and risks of language-generating AI. *Nature* 2021; 591: 22–25.
10. Korngiebel DM and Mooney SD. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery. *Npj Digit Med* 2021; 4: 1–3.
11. Regier DA, Narrow WE, Clarke DE, et al. DSM-5 field trials in the United States and Canada, part II: Test–retest reliability of selected categorical diagnoses. *Am J Psychiatry* 2013; 170: 59–70.
12. FitzGerald C and Hurst S. Implicit bias in healthcare professionals: A systematic review. *BMC Med Ethics* 2017; 18: 19.
13. Halladay AK, Bishop S, Constantino JN, et al. Sex and gender differences in autism spectrum disorder: Summarizing evidence gaps and identifying emerging areas of priority. *Mol Autism* 2015; 6: 36.
14. Skodol AE and Bender DS. Why are women diagnosed borderline more than men? *Psychiatr Q* 2003; 74: 349–360.
15. Parikh RB, Teeple S and Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019; 322: 2377–2378.
16. Sezgin E, Sirrianni J and Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: Outlook of generative pretrained transformer 3 (GPT-3) as a service model. *JMIR Med Inform* 2022; 10: e32875.
17. Gianfrancesco MA, Tamang S, Yazdany J, et al. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; 178: 1544–1547.
18. National Board of Medical Examiners. *Subject examinations: Content, outlines, and sample items*. National Board of Medical Examiners; 2021.
19. American Psychiatric Association (eds). *Diagnostic and statistical manual of mental disorders: DSM-5*. 5th ed. Washington, D.C: American Psychiatric Association; 2013.
20. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. ArXiv200514165 Cs [Internet]. 2020 Jul 22 [cited 2022 Feb 1]. Available from: http://arxiv.org/abs/2005.14165
21. Open AI. OpenAI API [Internet]. OpenAI. 2020 [cited 2022 Mar 25]. Available from: https://openai.com/blog/openai-api/
22. Kuhn M. The caret package [Internet]. [cited 2022 Mar 28]. Available from: https://topepo.github.io/caret/
23. Bates D, Maechler M, Bolker B, et al. lme4: Linear mixed-effects models using "Eigen" and S4 [Internet]. 2022 [cited 2022 Mar 31]. Available from: https://CRAN.R-project.org/package=lme4
24. Rai A. Explainable AI: From black box to glass box. *J Acad Mark Sci* 2020; 48: 137–141.
25. Schwartz RC and Blankenship DM. Racial disparities in psychotic disorder diagnosis: A review of empirical literature. *World J Psychiatry* 2014; 4: 133–140.
26. U.S. DHHS. 2020 National survey of drug use and health (NSDUH) releases | CBHSQ Data [Internet]. Substance Abuse and Mental Health Services Administration. 2020 [cited 2022 Apr 3]. Available from: https://www.samhsa.gov/data/release/2020-national-survey-drug-use-and-health-nsduh-releases
27. Skogli EW, Teicher MH, Andersen PN, et al. ADHD In girls and boys – gender differences in co-existing symptoms and executive function measures. *BMC Psychiatry*. 2013;13:298.
28. Solaiman I and Dennison C. Process for adapting language models to society (PALMS) with values-targeted datasets. 43.