BMJ Mental Health DIGITAL MENTAL HEALTH

Emulated trial of artificial intelligence use and subsequent depressive outcomes in a survey of US adults

Roy H Perlis [1,2] Faith M Gunning,[3] Ata Uslu,[4,5] Mauricio Santillana [4,5] Matthew A Baum,[6] James N Druckman [7] Katherine Ognyanova [8] David Lazer [4,5]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjment-2026-302609).

[1]Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA
[2]Department of Psychaitry, Harvard Medical School, Boston, Massachusetts, USA
[3]Department of Psychiatry, Weill Cornell Medicine, New York, New York, USA
[4]Network Science Institute, Northeastern University, Boston, Massachusetts, USA
[5]Institute for Quantitative Social Science, Harvard University, Boston, Massachusetts, USA
[6]John F. Kennedy School of Government and Department of Government, Harvard University, Cambridge, Massachusetts, USA
[7]Department of Political Science, University of Rochester, Rochester, New York, USA
[8]Department of Communication, Rutgers University New Brunswick, New Brunswick, New Jersey, USA

Correspondence to
Professor Roy H Perlis; RPERLIS@mgh.harvard.edu

Received 13 February 2026
Accepted 26 April 2026

To cite: Perlis RH, Gunning FM, Uslu A, *et al*. *BMJ Ment Health* 2026;**29**:1–8.

ABSTRACT

Background Generative artificial intelligence (AI) use has been suggested to have adverse mental health consequences but a causal relationship has not been examined.

Objective To simulate a randomised controlled trial of AI use in a work, school or personal context by applying target trial emulation to multiple waves of data from a nationally representative survey.

Methods We conducted a target trial emulation using non-probability survey data from three waves of a nationally representative survey conducted between 18 June 2024 and 8 January 2025. Participants aged ≥18 years reported generative AI use frequency at baseline. High-frequency use was defined as multiple times per week or more. The primary outcome was depressive symptom severity measured using the Patient Health Questionnaire 9-item (PHQ-9) at follow-up. Generalised causal forests assessed heterogeneity of treatment effects.

Findings Among 19 099 participants assessed at baseline, 2862 (15.0%) reported AI use at least multiple times per week. A subset of 3109 (16.3%) returned for follow-up. In the primary weighted analysis, high-frequency use was not significantly associated with change in PHQ-9 score at follow-up (mean difference −0.18, 95% CI −0.94 to 0.59; p=0.65). Multiple sensitivity analyses using alternate outcome definitions also did not identify significant causal effects. Generalised causal forests yielded no significant evidence of heterogeneity of effect (p=0.81).

Conclusions In an emulated randomised trial among US adults, generative AI use was not associated with subsequent depressive symptoms. This result does not support the premise that AI use causes greater depressive symptoms, although adverse outcomes among vulnerable individuals cannot be excluded.

Clinical implications AI use is unlikely to cause increased depressive symptoms among most US adults. Continued monitoring should clarify potential risks among vulnerable populations.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Prior cross-sectional studies have reported associations between generative artificial intelligence (AI) use and depressive symptoms, but these designs cannot establish causality and may reflect confounding or reverse causation.

WHAT THIS STUDY ADDS

⇒ In a target trial emulation using longitudinal survey data from US adults, regular AI use was not associated with subsequent depressive symptoms after adjusting for baseline differences and attrition, with no evidence of clinically meaningful effects or subgroup heterogeneity.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These findings do not support a causal relationship between AI use and depression at the population level, suggesting that public health concerns should focus on identifying vulnerable subgroups and evaluating longer-term effects rather than assuming broad population risk.

BACKGROUND

Generative artificial intelligence (AI) use has been rapidly embraced by the general public for work, school and personal use. Anecdotal evidence suggests that such use may contribute to psychosis,[1] suicidality[2] or other adverse mental health outcomes in vulnerable individuals. Other forms of online behaviour like social media use[3] may similarly contribute to depressive and anxious symptoms by diminishing direct human contact. More generally, the rapid transformation of work and school interactions driven by AI use could also negatively impact well-being.[4]

At least three studies, adopting different methodologies, have examined the association between AI use and depressive symptoms. One preprinted study of ChatGPT conversations found reduced social engagement, and symptoms labelled as dependence, among those with large amounts of use on a daily basis.[5] Greater daily use was also associated with depressive symptoms in a study of college students.[6] Finally, a subsequent large survey study of US adults in spring 2025 likewise suggested

modest but statistically significant associations between greater AI use and depressive symptoms.[7] A key limitation in all of these studies is their cross-sectional design, which precludes conclusions about causality of the relationship between generative AI use and mental health.

The most robust means of investigating causation requires randomised controlled trials, which may not always be feasible: exposure to suspected risk may not be ethical, or may be widespread enough that a true control group is difficult to achieve. Alternative randomised designs, such as randomised discontinuation, may represent a viable strategy but do not directly test the effects of an exposure among those not previously exposed.

In the case of generative AI, rapid uptake of this technology renders a randomised trial challenging to conduct as use has become so widespread. Target trial emulation provides an alternative approach to estimate causal effects using observational data.

OBJECTIVE

This study aimed to examine whether regular generative AI use is associated with subsequent depressive symptoms among US adults. Using longitudinal data from multiple waves of a nationally representative internet survey, we applied a target trial emulation framework to estimate the causal effect of high-frequency AI use on depressive symptom severity. This approach[8] matches exposed and unexposed individuals and applies stringent criteria to define and emulate measurement of exposure and outcome that would be expected in a randomised controlled trial.

METHODS

Study design

We used data from the Civic Health and Institutions Project (CHIP-50),[9] a consortium of academic sites which has conducted non-probability[10] internet surveys via a commercial survey panel aggregator (PureSpectrum) since early 2020. Such surveys do not sample purely at random, because participants opt in to participation; they therefore rely on quotas and reweighting to maximise representativeness.[11–13] For these analyses, we used survey wave 32 (from 18 June to 21 July 2024), which included questions about AI use, as the baseline visit, and wave 33 (from 30 August to 12 October 2024) and wave 34 (from 9 November 2024 to 8 January 2025) as follow-up visits, with only the first follow-up visit included.

Survey participants were required to be age 18 or older and to reside in one of the 50 US states or the District of Columbia. They were provided the option to participate in a general opinion survey. Separate surveys were conducted by state, with quotas for age, gender, and racial and ethnic group participation. Surveys included attention checks requiring respondents to provide a specific response to avoid automated, fraudulent or inattentive participants.

The survey was conducted and presented in accordance with the AAPOR transparency initiative compliance checklist.[14]

Measures

Exposure: use of AI by self-report

We asked all participants in Wave 32, 'How often do you use the following technologies or products?—artificial intelligence (AI)'; options included never, once or twice, about once a month, about once a week, multiple times a week, every day, and multiple times a day. Since we expected that some participants might recognise alternative terms (generative AI, large language models), we asked the same question with that phrasing; as results were nearly identical in our prior work,[7] we used only the primary question here. For purposes of developing weights, as described below, we treated this variable ordinally; for purposes of trial emulation, it was dichotomised to multiple times per week versus once a week or less. To determine the nature of use, participants were also asked, 'For what purposes do you use the following technologies or products? (Please select all that apply)—artificial intelligence' and could check 'personal use', 'for work' and/or 'for school'.

Covariates: sociodemographic variables and use of social media

Sociodemographic characteristics were collected by self-report and categorised as in prior survey reports[3 15–17] from this project. Race and ethnicity were selected from a list including African American or Black, Asian American, Hispanic, Native American, Pacific Islander, white or other, with the opportunity to provide a free text self-description. These measures were collected to allow us to confirm representativeness of the US population and are reported in accordance with guidelines for use of such data.[18] Consistent with our prior work, as some individual groups were small, we collapsed Native American, Pacific Islander, individuals indicating multiple racial identities and other into a single category to facilitate analysis. For gender, individuals could identify as male, female, genderqueer or other with free-text entry; those in the latter two categories were combined into 'Another gender' to facilitate analysis given the smaller sample size. Household socioeconomic status in terms of annual household income was categorised as <$25k, $25k–<$50k, $50k–<$75k, $75–<$100k or $100k+ per year. Educational status was determined by asking participants to select from their highest level of formal education from a list including some high school or less, high school graduate, some college, undergraduate degree or graduate degree. Urban/rural status was determined based on zip code using the USDA's Rural-Urban Commuting Area Codes.[19]

We also asked participants how often they use specific social media platforms, and how often they post to specific social media platforms. For these analyses, we examined maximum use across all platforms and maximum posting across all platforms. For use, participants were asked if they use a given platform and then 'How often do you use the following social media platforms? —{platform_name}'. For individuals who reported using a given platform, frequency of use on a 6-point scale ranged from 'less than once a week' to 'most of the day'. (For analysis, this was converted to a 7-point scale by adding 'never' for individuals who do not use the platform at all). Respondents were also asked, 'How often do you post anything on the following platforms?' Frequency of posting on a 6-point scale ranged from 'never' to 'multiple times a day'.

Participants also completed the University of California, Los Angeles (UCLA) 3-item loneliness scale,[20 21] which asks, 'How often do you feel that you lack companionship?', 'How often do you feel left out?' and 'How often do you feel isolated from others?'. The three response categories are 'Hardly ever' for a score of 1, 'Some of the time' for a score of 2 or 'Often' for a score of 3; these are summed to yield a composite score of 3–9.

Outcome: depressive and anxious symptoms

Depressive symptom severity was measured with the 9-item Patient Health Questionnaire (PHQ-9),[22 23] a standard screening scale encompassing diagnostic criteria for major depressive disorder in the Diagnostic and Statistical Manual of Mental

Disorders, Fifth Edition (DSM-5); respondents rate each symptom in terms of frequency over the prior 2 weeks on a 0–3 Likert-type scale (0=not at all, 3=nearly every day), and these values are summed to yield a total score from 0 to 27. Scoring 10 or greater indicates at least moderate depressive symptoms and is typically considered a threshold for treatment referral.[22 23] We similarly examined anxiety symptoms using the two-item Generalized Anxiety Disorder scale (GAD-2), which uses the same timeframe and anchor points.[24 25]

Statistical analysis

The protocol for the emulated target trial was specified as follows:

Eligibility: Participants who completed wave 32 of the survey and had complete data on AI usage and baseline covariates.

Intervention: We compared two strategies: (1) 'High-Frequency AI Use' (defined as using AI tools multiple times per week or more frequently) versus (2) 'Low/No AI Use' (defined as using AI tools once a week or less frequently).

Treatment assignment: Participants were classified into treatment groups based on their self-reported usage at wave 32 (baseline).

Follow-up and outcome: The primary outcome was the severity of depressive symptoms, measured via total PHQ-9 score. Outcomes were ascertained at the first available follow-up wave, either wave 33 or wave 34. A secondary analysis examined severity of anxious symptoms, measured via total GAD-2 score.

Covariates: To account for confounding, we extracted baseline characteristics measured at wave 32. These included sociodemographic factors (age, gender, race and ethnicity, educational attainment, household income, urbanicity), baseline depressive symptoms (wave 32 total PHQ-9 score) and behavioural proxies for technology adoption (frequency of social media use and social media posting).

Primary analysis used an inverse probability weighting (IPW) approach to adjust for both confounding by indication and selection bias due to loss to follow-up (attrition). For propensity score weighting for treatment, we estimated the probability of high-frequency AI use using a multivariable logistic regression model conditioned on all baseline covariates. Stabilised weights for treatment were calculated as the ratio of the marginal probability of the observed treatment to the conditional probability. We also calculated inverse probability of censoring weighting (IPCW); given the substantial expected attrition between baseline and follow-up, we modelled the probability of remaining in the study (ie, being 'uncensored'). A logistic regression model predicted study retention based on both the treatment group (AI use) and all baseline covariates. We calculated stabilised censoring weights to up-weight participants who remained in the study but resembled those who dropped out, under the assumption that data were missing at random *conditional* on the observed covariates. The final participant weight was therefore the product of the treatment weight and the censoring weight. We estimated the average treatment effect using a weighted generalised linear model with a Gaussian distribution and identity link.

Robust variance estimators (sandwich estimators) were used to calculate 95% CIs to account for the weighting induced correlation. Standardised mean differences (SMDs) were calculated to assess covariate balance, with an SMD <0.1 considered indicative of ignorable imbalance.

Sensitivity analyses are summarised in online supplemental methods. We also explored heterogeneity of treatment effects using generalised causal forests (see online supplemental methods).[26–29]

All analyses used R V.4.4.2. Key packages included *ipw* for weight generation, *survey* for weighted modelling, *sandwich* for robust standard errors and *cobalt* for assessing covariate balance. Imputation used *mice*. For all analyses, two-tailed p<0.05 was considered to represent statistical significance.

FINDINGS

Study population

A total of 19 099 participants met the eligibility criteria at wave 32 (baseline assessment); their characteristics are summarised in online supplemental table 1. Of these, 2862 (15.0%) reported AI use at least multiple times per week, while 16 237 (85.0%) reported less frequent or no use. By the follow-up assessment (wave 33/34), 15 990 participants (83.7%) were lost to follow-up, leaving a final analytical sample of 3109 (16.3%). Characteristics of this sample are summarised in table 1. Figure 1 illustrates the derivation of the study cohort and attrition.

Covariate balance and attrition

In the propensity score model estimating the likelihood of high-frequency AI use (online supplemental table 2; online supplemental figure 5), features associated with such use included younger age (eg, age 65+ vs 18–24: OR=0.46, 95% CI 0.38 to 0.56) and having higher baseline PHQ-9 scores (OR=1.01 per point increase, 95% CI 1.01 to 1.02). Higher education (eg, graduate degree vs some high school or less: OR=2.08, 95% CI 1.55 to 2.83) and higher household income (eg, $100k+ vs under $25k: OR=1.53, 95% CI 1.33 to 1.77) were also associated with a greater likelihood of use.

Attrition analysis revealed that dropout was not random. Modelling study retention (ie, the likelihood of returning for follow-up), features associated with retention included older age (eg, age 65+ vs 18–24: OR=6.27, 95% CI 4.67 to 8.60) and lower baseline PHQ-9 scores (OR=0.98 per point increase, 95% CI 0.97 to 0.99). Retention also differed by race: compared with participants identifying as Asian (reference), White participants were significantly less likely to return (OR=0.71, 95% CI 0.60 to 0.85). Those individuals using AI multiple times a week were more likely to return compared with those who stated they never used AI (OR=1.27, 95% CI 1.09 to 1.49) (online supplemental table 3; online supplemental figure 6).

After applying the combined stabilised weights, covariate balance improved substantially. All SMDs between the exposure groups were reduced to below 0.10 (41 out of 41 covariates balanced), indicating successful emulation of randomisation (figure 2).

Primary analysis: AI use and depressive symptoms

In the primary weighted analysis, applying inverse probability weighting for both treatment and censoring, higher-frequency AI use was not significantly associated with change in PHQ-9 score at follow-up (mean difference −0.18, 95% CI −0.94 to 0.59; p=0.65) compared with low/no use (figure 3). Omitting weights for censoring (ie, assuming censoring from attrition was uninformative) yielded a 0.00 point change (95% CI −0.58 to 0.58; p=1.00). Post hoc analysis adding UCLA-3 to models as a measure of loneliness did not yield meaningful differences, with change in PHQ-9 of −0.15 (95% CI −0.92 to 0.61; p=0.69). Likewise, incorporating terms for follow-up wave yielded similar results, with change in PHQ-9 of −0.15 (95% CI −0.91 to 0.60; p=0.69) (online supplemental table 7).

Table 1 Characteristics of analytical sample, including sociodemographic and other features

Characteristic	N	Overall n=3109*	Low/no AI Use n=2674*	High-frequency AI use n=435*	P value†
Age (years)	3109	59.9 (15.0)	60.8 (14.7)	53.9 (15.5)	**<0.001**
Gender	3109				**<0.001**
Another gender		13 (0.4%)	9 (0.3%)	4 (0.9%)	
Female		1602 (52%)	1417 (53%)	185 (43%)	
Male		1494 (48%)	1248 (47%)	246 (57%)	
Race	3109				**<0.001**
Asian		191 (6.1%)	155 (5.8%)	36 (8.3%)	
Black		336 (11%)	260 (9.7%)	76 (17%)	
Other		123 (4.0%)	101 (3.8%)	22 (5.1%)	
White		2459 (79%)	2158 (81%)	301 (69%)	
Ethnicity	3109				**<0.001**
Non-Hispanic		2874 (92%)	2490 (93%)	384 (88%)	
Hispanic/Latino		235 (7.6%)	184 (6.9%)	51 (12%)	
Education	3109				**0.006**
Some high school or less		52 (1.7%)	44 (1.6%)	8 (1.8%)	
High school graduate		643 (21%)	580 (22%)	63 (14%)	
Some college		767 (25%)	664 (25%)	103 (24%)	
College degree		1211 (39%)	1016 (38%)	195 (45%)	
Graduate degree		436 (14%)	370 (14%)	66 (15%)	
Household income	3109				**0.001**
Under 25k		602 (19%)	532 (20%)	70 (16%)	
25k to under 50k		928 (30%)	817 (31%)	111 (26%)	
50k to under 75k		673 (22%)	580 (22%)	93 (21%)	
75k to under 100k		373 (12%)	312 (12%)	61 (14%)	
100k and over		533 (17%)	433 (16%)	100 (23%)	
Urbanicity	3109				**<0.001**
Rural		490 (16%)	441 (16%)	49 (11%)	
Suburban		1816 (58%)	1571 (59%)	245 (56%)	
Urban		803 (26%)	662 (25%)	141 (32%)	
PHQ-9 score (baseline)	3109	4.2 (5.5)	4.1 (5.4)	5.0 (5.9)	**0.006**
GAD-2 score (baseline)	3108	1.0 (1.6)	1.0 (1.6)	1.2 (1.7)	**0.007**
University of California, Los Angeles Loneliness Scale-3 score (baseline)	3109	4.6 (1.9)	4.6 (1.9)	4.7 (2.0)	0.2
Social media use frequency	3109				**<0.001**
Never		504 (16%)	469 (18%)	35 (8.0%)	
<1x/week		131 (4.2%)	118 (4.4%)	13 (3.0%)	
1x/week		156 (5.0%)	141 (5.3%)	15 (3.4%)	
Several times/week		253 (8.1%)	222 (8.3%)	31 (7.1%)	
1 x/day		569 (18%)	497 (19%)	72 (17%)	
Several times/day		1073 (35%)	901 (34%)	172 (40%)	
Most of the day		423 (14%)	326 (12%)	97 (22%)	
Social media posting frequency	3109				**<0.001**
Never		1017 (33%)	928 (35%)	89 (20%)	
<1x/month		621 (20%)	559 (21%)	62 (14%)	
1x/month		275 (8.8%)	229 (8.6%)	46 (11%)	
1x/week		527 (17%)	437 (16%)	90 (21%)	
1x/day		364 (12%)	297 (11%)	67 (15%)	
Multiple times/day		305 (9.8%)	224 (8.4%)	81 (19%)	
PHQ-9 score (follow-up)	3109	4.8 (5.7)	4.8 (5.6)	5.4 (5.7)	**0.023**
GAD-2 score (follow-up)	3108				**0.029**
0		1853 (60%)	1613 (60%)	240 (55%)	
1		364 (12%)	317 (12%)	47 (11%)	
2		475 (15%)	401 (15%)	74 (17%)	
3		116 (3.7%)	96 (3.6%)	20 (4.6%)	
4		125 (4.0%)	99 (3.7%)	26 (6.0%)	
5		64 (2.1%)	49 (1.8%)	15 (3.4%)	

Continued

Table 1 Continued

Characteristic	N	Overall n=3109*	Low/no AI Use n=2674*	High-frequency AI use n=435*	P value†
6		111 (3.6%)	98 (3.7%)	13 (3.0%)	
Follow-up wave	3109				0.7
Wave 33		2026 (65%)	1746 (65%)	280 (64%)	
Wave 34		1083 (35%)	928 (35%)	155 (36%)	
Days to follow-up	3109	100.3 (39.1)	100.1 (39.0)	101.3 (39.5)	0.6

Values reported as mean (SD) for continuous and n (%) for categorical measures. P values from χ^2 tests (categorical) or t tests (continuous). Bolded p values indicate statistical significance (p<0.05).
*Continuous measures reported as mean (SD); categorical measures as n (%).
†Welch's two sample t test; Pearson's χ^2 test.
AI, artificial intelligence; GAD-2, two-item Generalized Anxiety Disorder scale; PHQ-9, Patient Health Questionnaire 9-item.

For context, the unadjusted analysis showed a difference of +0.67 points (95% CI 0.10 to 1.25; p=0.02). After adjusting for baseline PHQ-9 alone, the effect was attenuated to +0.02 points (95% CI −0.36 to 0.40; p=0.91).

We also examined change in GAD-2 as a secondary measure, yielding a pattern of results similar to that observed for PHQ-9. Applying inverse probability weighting for both treatment and censoring, the average treatment effect for higher-frequency AI use was a 0.03-point increase in total GAD-2 scores at follow-up (95% CI −0.20 to 0.24; p=0.83) compared with low/no use (online supplemental figure 4; online supplemental table 8). Unadjusted analysis showed a difference of 0.19 points (95% CI 0.03 to 0.35; p=0.02), while adjustment for baseline GAD-2 alone yielded 0.02 points (95% CI −0.09 to 0.14; p=0.69).

Sensitivity analyses

We next examined whether a higher threshold for exposure—that is, requiring daily rather than multiple times per week use of generative AI—would yield meaningfully different results. Daily AI use was associated with a 0.60 point decrease in PHQ-9 (ie, an average treatment effect of −0.60) at follow-up (95% CI −1.59 to 0.40; p=0.24) compared with less frequent or no use (figure 3; for propensity score model, see online supplemental table 5; online supplemental figure 9); covariate balance shown in online supplemental figure 10. With an even higher threshold for frequent use, namely, multiple times per day, we identified a 1.03-point decrease in PHQ-9 at follow-up which was not statistically significant (95% CI −2.55 to 0.48; p=0.18), compared with less frequent or no use (figure 3; for propensity score model, see online supplemental table 6; online supplemental figure 11; covariate balance shown in online supplemental figure 12).

We then defined the exposure as being limited to personal use of generative AI at least multiple times per week, and repeated target trial emulation. Here, personal use was associated with a 0.28 point decrease in PHQ-9 (95% CI −1.10 to 0.54; p=0.51) compared with less frequent use (figure 3; for propensity score model see online supplemental table 4; online supplemental figure 7; covariate balance shown in online supplemental figure 8).

We further conducted tipping point analysis for attrition (online supplemental figure 1). Under extreme assumptions—where dropouts among AI users had PHQ-9 scores 10 points lower (δ=−10) than observed participants, or 10 points higher (δ=+10), effect estimates ranged from 0.53 to 0.76 points. No estimates approached the prespecified minimal clinically important difference of 3 points. We also conducted quantitative bias analysis to determine whether unmeasured confounding could have masked a clinically meaningful effect, again defined as a between-group difference of 3 points on the PHQ-9 (online supplemental figure 2). To mask such an effect, we estimate that a confounder (beyond those adjusted for in our models) would need to be associated with an ~1.75 fold increase in the likelihood of AI use and independently associated with the outcome by a factor of 2.0, or approximately 2.4 points on PHQ-9.

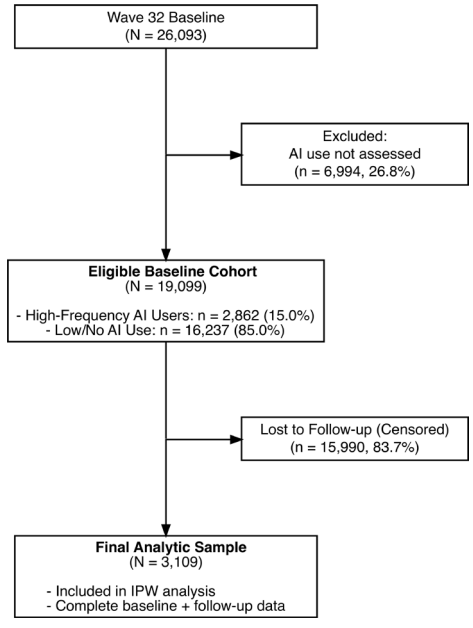

Figure 1 CONSORT (Consolidated Standards of Reporting Trials) diagram illustrating study cohort derivation and attrition from wave 32 baseline to wave 33/34 follow-up. AI, artificial intelligence; IPW, inverse probability weighting.

Examination of heterogeneity

Finally, in a GCF analysis examining the primary treatment, defined as any AI use at least multiple times per week, a test for heterogeneity of treatment effect across all of the sociodemographic and clinical variables at baseline yielded p=0.81. Increasing the threshold for treatment to daily AI use resulted in heterogeneity p=0.7, while limiting the exposure to personal AI use yielded p=0.54. In other words, none of these analyses were consistent with significant heterogeneity across subgroups. A post hoc analysis examining exposure-by-race interactions in the primary analysis was not statistically significant (p=0.13); for descriptive purposes, online supplemental figure 3 plots effects stratified by race and ethnicity.

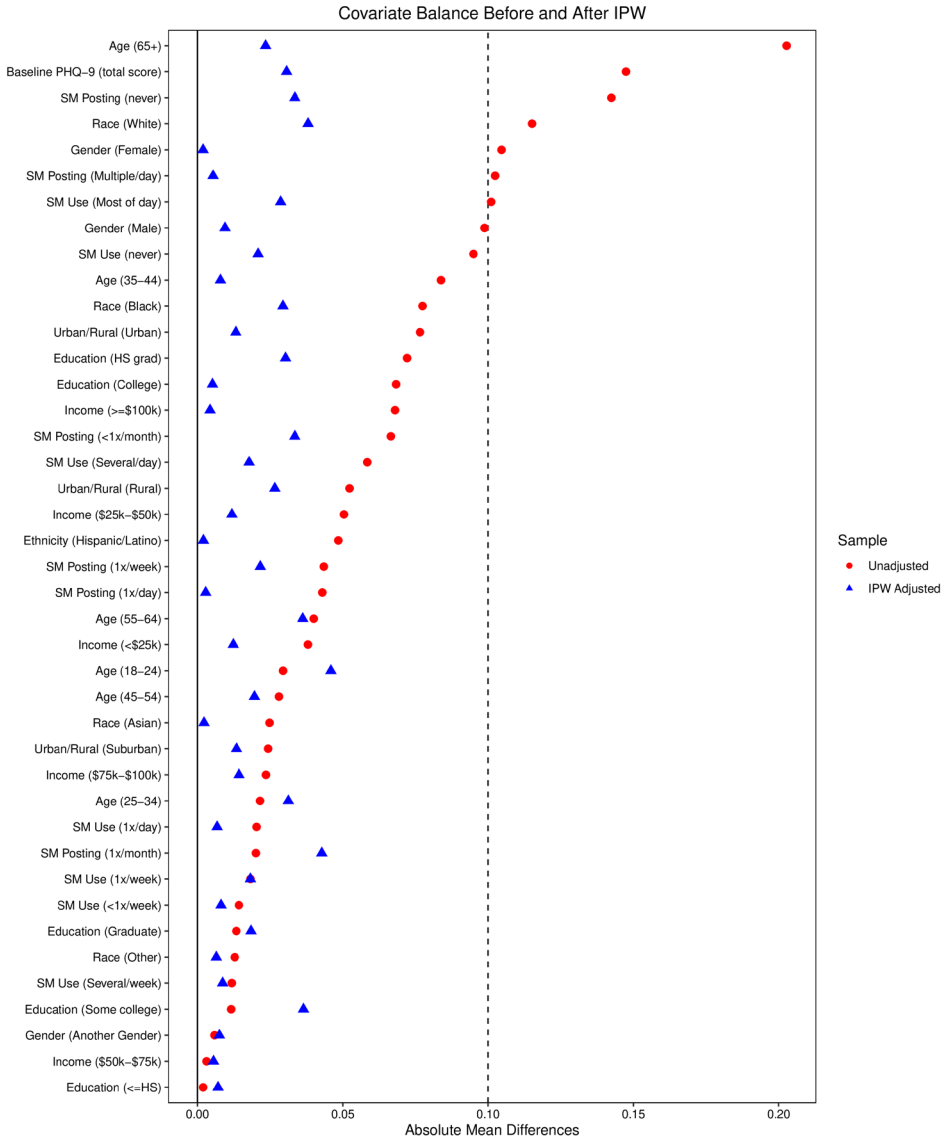

Figure 2 Covariate balance in terms of standardised mean (SM) differences in the matched exposed/unexposed groups after stabilised weighting. IPW, inverse probability weighting.

DISCUSSION

In this target trial emulation study of 19 099 participants, we found no statistically significant association between high-frequency AI use and subsequent depressive symptoms after adjustment for potential confounding variables and attrition bias using inverse probability weighting. Very modest evidence of an association observed in unadjusted analyses was entirely accounted for by baseline differences between AI users and non-users, and by differential loss to follow-up. This lack of association between generative AI use and depressive symptoms remained robust to sensitivity analyses examining alternative definitions of exposure.

A prior cross-sectional analysis using large-scale survey data, conducted in spring 2025, suggested modest association between AI use and greater depressive symptoms.[7] Two prior studies also suggested the potential for depressive symptom worsening among some individuals, including a survey of college students conducted in China[6] and a preprint examining ChatGPT conversations.[5] In all three of these studies, greater magnitude of use was associated with greater risk for depressive symptoms. These studies align with anecdotal evidence suggesting that chatbot use may reinforce suicidality[2] or delusional thinking.[1]

However, cross-sectional design in previous studies precluded any examination of causation; in particular, it could not exclude the possibility of reverse causation (ie, greater AI use caused by depressive symptoms) or confounding (ie, individual characteristics independently associated with both AI use and depressive symptoms). Relying instead on target trial emulation,[8] we did not identify causal effects for AI use on depressive symptoms, with sensitivity analyses supporting the robustness of this result, suggesting that prior findings may well reflect confounding or other sources of bias.

We also sought to investigate whether there might be particular sociodemographically defined subgroups at greater risk for depressive symptom increase with AI use. That is, even absent a detectable effect in the population on average, it is possible that subgroups exist where AI use is more harmful. In lieu of testing each subgroup individually, we applied a machine learning-based approach that seeks to identify heterogeneity of treatment effects as an omnibus test, to avoid risk for type 1 error. This

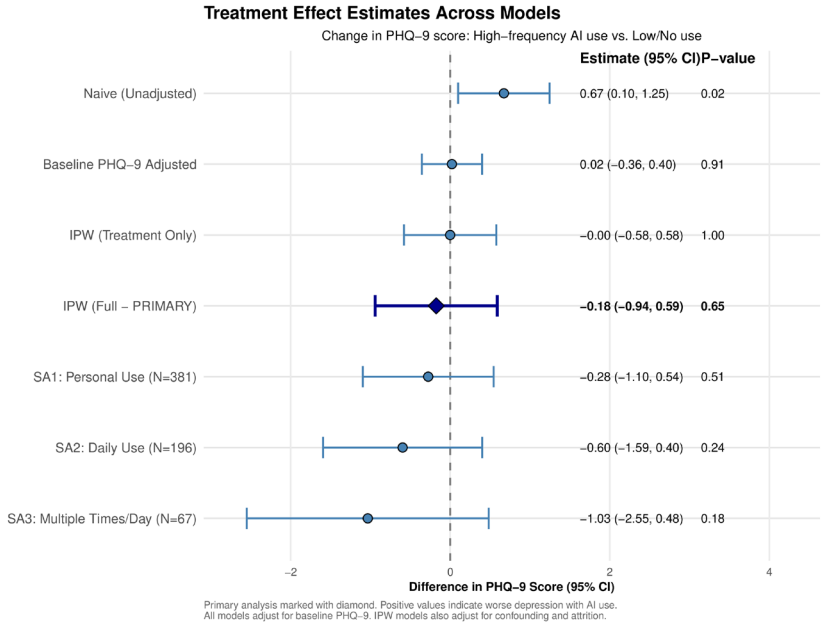

Figure 3 Point estimates and CIs for effect sizes in emulated target trial, across designs and including sensitivity analysis. PHQ-9, IPW, inverse probability weighting; PHQ-9, Patient Health Questionnaire 9-item.

test for heterogeneity did not identify significant variability by groups, either for the primary exposure definition or for the two secondary definitions.

Limitations

In seeking to model a randomised controlled trial, we addressed multiple key challenges to validity. First, we cannot distinguish duration of exposure, namely, whether individuals only recently began using AI or had been using it chronically. We elected to use the earliest large-scale measurement of AI use in this survey, which occurred in June/July 2024, during a period of rapid increase in application of large language models when use of generative AI was not yet widespread. However, among individuals who had already been using such tools for more than a year, it is possible that any negative effects had already accrued, biasing us toward the null hypothesis of no effect. In other words, if AI has an immediate negative impact which is sustained, we would not detect it with this approach. To address this possibility, we included multiple exposure definitions (including amount of use and type of use), but we could not directly examine duration of use or specific use cases beyond personal vs work or education. We also cannot estimate the potential effects of chronic use, as our results are limited in follow-up duration. Second, as the survey is not a fixed panel but rather allows participants to opt in to any given wave, our attrition rate is substantial by comparison with many clinical trials. We addressed this concern by explicitly modelling attrition—that is, examining propensity to drop out, as well as treatment propensity—which should reduce bias arising from nonrandom missingness. Third, as with any nonrandomised study, emulating a trial requires that we adequately model differences between treated and untreated individuals. Our quantitative bias analysis suggests that a substantial unmeasured confounder would be required to explain our null finding in the presence of a true 3-point difference—such a confounder would need to double the likelihood of AI use and be associated with a nearly 5-point difference in PHQ-9. Finally, our approach cannot exclude modest population differences or effects that are observed only in a small subset of individuals. A particular priority may be understanding risk among individuals with greater liability for delusions or suicidality, for example, see previous work[30]—but alternate study designs will likely be required.

CONCLUSIONS

Our emulated clinical trial, drawing on large-scale longitudinal survey data, does not identify evidence of a causal link between regular AI use and depressive symptoms. While this approach cannot rule out substantial effects in a very small subset of individuals, it suggests that prior cross-sectional findings may have arisen from confounding and that population-wide effects are likely to be modest on average.

Contributors RHP conceptualised the study, conducted formal analyses, drafted the manuscript and is the guarantor. FMG contributed to analytic design and revised manuscript. AU, MS, MAB, JND, KO and DL contributed to data collection and revised manuscript. RHP used Anthropic Claude (Sonnet 4.5) and ChatGPT 5.2 (thinking) to debug analytic and figure generation code where required; he takes full responsibility for all such code and its outputs.

Funding This study was supported by the National Institute of Mental Health (RHP and DL, RF1MH132335; RHP, 1U01MH136059-01) and the National Science Foundation grants (KO, DL, JND and MAB SES-2029292, SES-2029792, SES2116465, SES-2116189). The sponsors did not have any role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; and decision to submit the manuscript for publication. The authors had the final responsibility for the decision to submit for publication. RHP had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Competing interests RHP has received consulting fees from Alkermes, Circular Genomics and Genomind. He holds equity in Circular Genomics. RHP is Editor in Chief of *JAMA+ AI*; RHP and FMG are paid Associate Editors for *JAMA Network Open*. The other authors report no disclosures.

Patient consent for publication Not applicable.

Ethics approval All participants provided written online consent before answering survey questions. The Harvard University Institutional Review Board determined the survey to be exempt under protocol # IRB20-0593. The survey was conducted and presented in accordance with the AAPOR transparency initiative compliance checklist. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The survey used for this study is available from the corresponding author for non-commercial use.

ORCID iDs
Roy H Perlis https://orcid.org/0000-0002-5862-6757
Mauricio Santillana https://orcid.org/0000-0002-4206-418X
James N Druckman https://orcid.org/0000-0002-1249-6790
Katherine Ognyanova https://orcid.org/0000-0003-3038-7077
David Lazer https://orcid.org/0000-0002-7991-9110

REFERENCES

1 I feel like i'm going crazy': chatgpt fuels delusional spirals - WSJ. Available: https://www.wsj.com/tech/ai/i-feel-like-im-going-crazy-chatgpt-fuels-delusional-spirals-ae5a51fc?gaa_at=eafs&gaa_n=ASWzDAj6SLCII03Jf2qKZDie_5vJ13jlEeBsZu1F3yGKgUsx1rTJ4Fc48lmfa0woukE%3D&gaa_ts=68ab0a85&gaa_sig=5sk0-dOKhxRtACJ8Kuzn4zUam_jPDyKVMohU1cLablCAJDmaEHa2BlJewylVIlDGOT5WHZNrKWoYTWrbMoPUWA%3D%3D [Accessed 24 Aug 2025].
2 A Teen Was Suicidal. ChatGPT was the friend he confided in. The New York Times. Available: https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html [Accessed 11 Sep 2025].
3 Perlis RH, Uslu A, Schulman J, *et al*. Irritability and Social Media Use in US Adults. *JAMA Netw Open* 2025;8:e2452807.
4 Perlis RH. Artificial Intelligence and the Potential Transformation of Mental Health. *JAMA Psychiatry* 2026;83:409–13.
5 MIT Media Lab. Investigating affective use and emotional wellbeing on chatgpt. Available: https://www.media.mit.edu/publications/investigating-affective-use-and-emotional-well-being-on-chatgpt/ [Accessed 11 Aug 2025].
6 Zhang X, Li Z, Zhang M, *et al*. Exploring artificial intelligence (AI) Chatbot usage behaviors and their association with mental health outcomes in Chinese university students. *J Affect Disord* 2025;380:394–400.
7 Perlis RH, Gunning FM, Uslu A, *et al*. Generative AI Use and Depressive Symptoms Among US Adults. *JAMA Psychiatry* 2026;9.
8 Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol* 2016;183:758–64.
9 The civic health and institutions project | chip50. Available: https://www.chip50.org/ [Accessed 11 Aug 2025].
10 Kennedy C, Caumont R. What we learned about online nonprobability polls. Pew Research Center; 2016. Available: https://www.pewresearch.org/fact-tank/2016/05/02/q-a-online-nonprobability-polls/
11 Radford J, Green J, Quintana A, *et al*. OSF Preprints; Evaluating the generalizability of the COVID States survey — a large-scale, non-probability survey, 2022. 10.31219/osf.io/cwkg7
12 Perlis RH, Simonson MD, Green J, *et al*. Prevalence of Firearm Ownership Among Individuals With Major Depressive Symptoms. *JAMA Netw Open* 2022;5:e223245.
13 Druckman JN, Ognyanova K, Safarpour A, *et al*. Representation in science and trust in scientists in the USA. *Nat Hum Behav* 2026;10:476–91.
14 Transparency initiative - aapor. 2022. Available: https://aapor.org/standards-and-ethics/transparency-initiative/ [Accessed 29 Mar 2026].
15 Perlis RH, Green J, Simonson M, *et al*. Association Between Social Media Use and Self-reported Symptoms of Depression in US Adults. *JAMA Netw Open* 2021;4:e2136113.
16 Perlis RH, Ognyanova K, Uslu A, *et al*. Trust in Physicians and Hospitals During the COVID-19 Pandemic in a 50-State Survey of US Adults. *JAMA Netw Open* 2024;7:e2424984.
17 Perlis RH, Lunz Trujillo K, Safarpour A, *et al*. Community Mobility and Depressive Symptoms During the COVID-19 Pandemic in the United States. *JAMA Netw Open* 2023;6:e2334945.
18 Flanagin A, Frey T, Christiansen SL, *et al*. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA* 2021;326:621–7.
19 Rural-urban commuting area codes | economic research service. Available: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes [Accessed 23 Dec 2025].
20 Russell DW. UCLA Loneliness Scale (Version 3): reliability, validity, and factor structure. *J Pers Assess* 1996;66:20–40.
21 Hughes ME, Waite LJ, Hawkley LC, *et al*. A Short Scale for Measuring Loneliness in Large Surveys. *Res Aging* 2004;26:655–72.
22 Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606–13.
23 Levis B, Benedetti A, Thombs BD, *et al*. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* 2019;365:l1476.
24 Spitzer RL, Kroenke K, Williams JBW, *et al*. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006;166:1092–7.
25 Plummer F, Manea L, Trepel D, *et al*. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. *Gen Hosp Psychiatry* 2016;39:24–31.
26 Chernozhukov V, Demirer M, Duflo E, *et al*. Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. *National Bureau of Economic Research* 2018. Available: https://www.nber.org/papers/w24678
27 Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Statist* 2019;47:1148–78.
28 Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J Am Stat Assoc* 2018;113:1228–42.
29 Tibshirani J, AtheyS, FriedbergR, *et al*. Grf: generalized random forests. 2023.
30 Hudon A, Stip E. Delusional Experiences Emerging From AI Chatbot Interactions or "AI Psychosis". *JMIR Ment Health* 2025;12:e85799.