AI and Behavioral Health: Data and Predictive Analytics

Shannon Wiltsey Stirman, PhD

Stanford University, Department of Psychiatry and Behavioral Sciences

National Dialogues for Behavioral Health | November 9, 2026

CREATE

THE CENTER FOR ADVANCING THERAPY WITH AI

Disclosures

- Research funding: National Institute of Mental Health, P50 MH139450 (CREATE Center)

Learning objectives

- Distinguish forms of AI from conventional statistical prediction
- Describe AI applied to health records and session text
- Summarize evidence for data-driven treatment personalization
- Explain validation and monitoring needs for predictive tools
- Describe consent and governance practices for clinical data

Forms of AI in behavioral health (1 of 2)

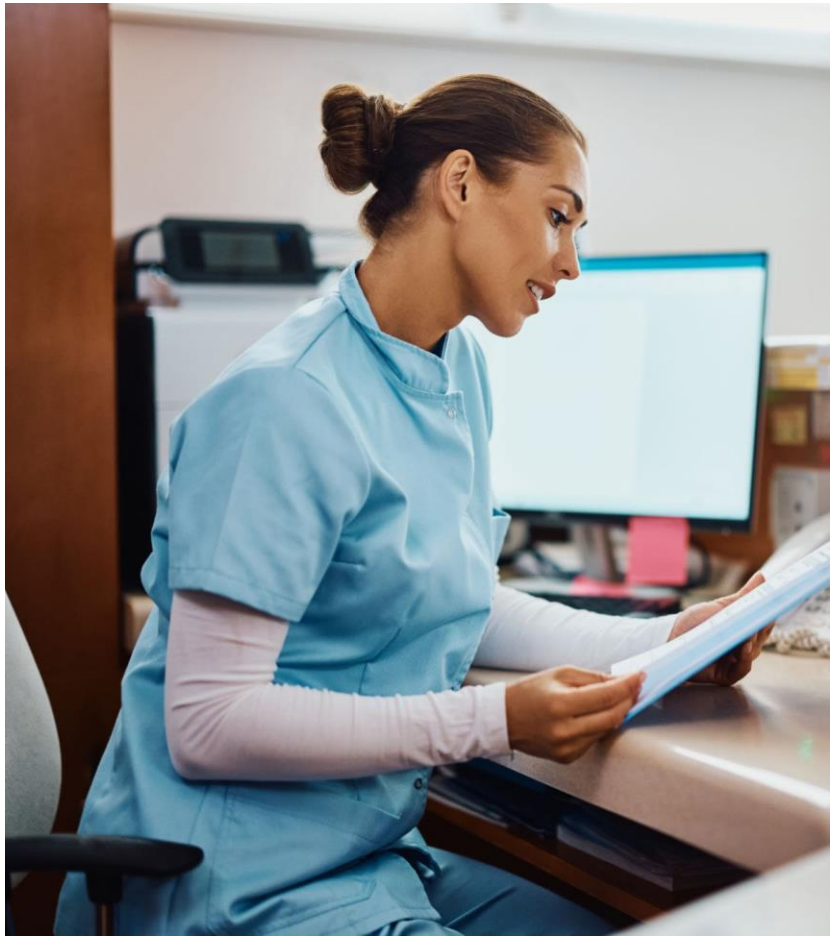
Approach	What it does	Examples
Statistical prediction (usually not called AI)	Estimates outcomes from selected variables	Outcome feedback, treatment selection indices
Machine learning	Finds patterns in large structured datasets	Stratified care, suicide risk models
Natural language processing	Classifies or summarizes text	Coding therapy notes, fidelity ratings

Forms of AI in behavioral health (2 of 2)

Approach	What it does	Examples
Speech recognition and deep learning	Converts speech to text	Transcription behind AI scribes
Large language models (generative AI)	Interprets and generates language	AI scribes, simulated patients, risk flags
Agentic AI (emerging)	Plans and carries out multi-step tasks	Care coordination prototypes

Data sources in routine care

- Symptom measures (PHQ-9, PCL-5)
- Structured health record data: visits, diagnoses, prescriptions
- Clinical notes
- Session audio and transcripts
- Client-facing apps and chatbots



Clients already use LLMs for mental health support

- 24% of 1,871 surveyed U.S. adults used LLMs for mental health
- Common uses: emotional support, coping strategies, someone to talk to
- These conversations are a new source of mental health data outside the clinic

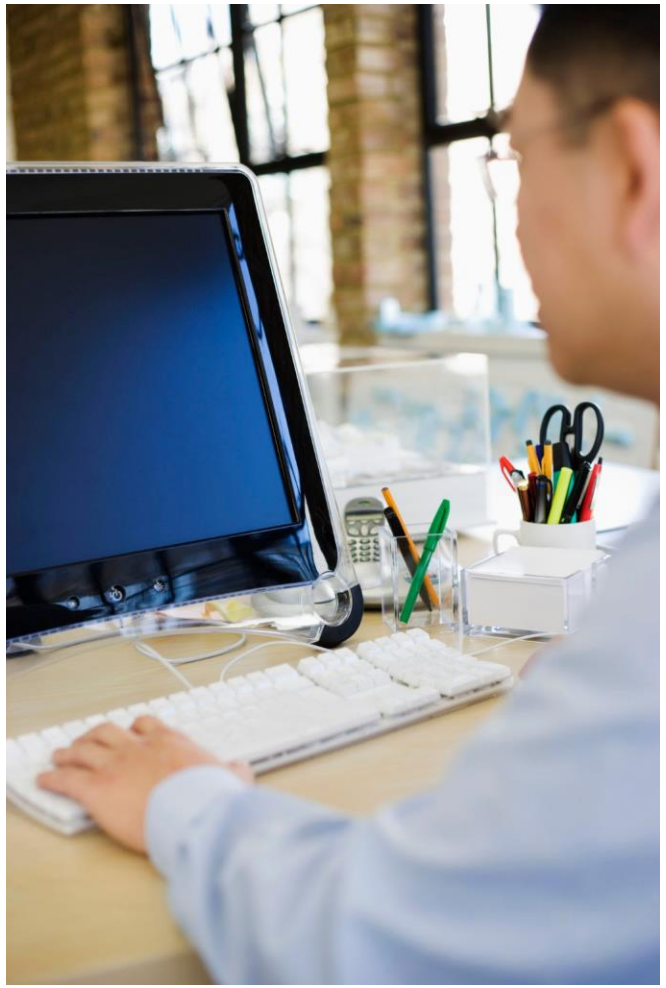


AI applied to clinical text

Measuring EBP delivery from VA notes

Method: natural language processing (AI) | Maguen et al., 2019

- Classified 8.1 million notes for 255,968 veterans with PTSD
- Accuracy .92 against human annotators
- 20.2% received at least one session of evidence-based psychotherapy



Effectiveness of EBPs in routine VA care

Method: NLP-derived data with causal statistics | Maguen et al., 2021

- Emulated target trials in 265,566 veterans with PTSD
- CPT and PE completers improved more than veterans in other psychotherapy
- Gains were smaller than in randomized trials

Psychotherapy notes before suicide deaths

Method: NLP topic modeling (AI) | Levis et al., 2026

- VA psychotherapy notes from veterans who died by suicide and matched controls
- Topics included risk, treatment planning, evaluation, psychosocial, medication
- Topic patterns differed by suicide risk tier

Automated coding of treatment fidelity

Method: natural language processing (AI) | Atkins et al., 2014

- Motivational interviewing fidelity coded from session transcripts
- Automated coding makes fidelity feedback possible at scale
- Commercial platforms such as Lyssn build on this work



CREATE examples: fidelity and risk detection

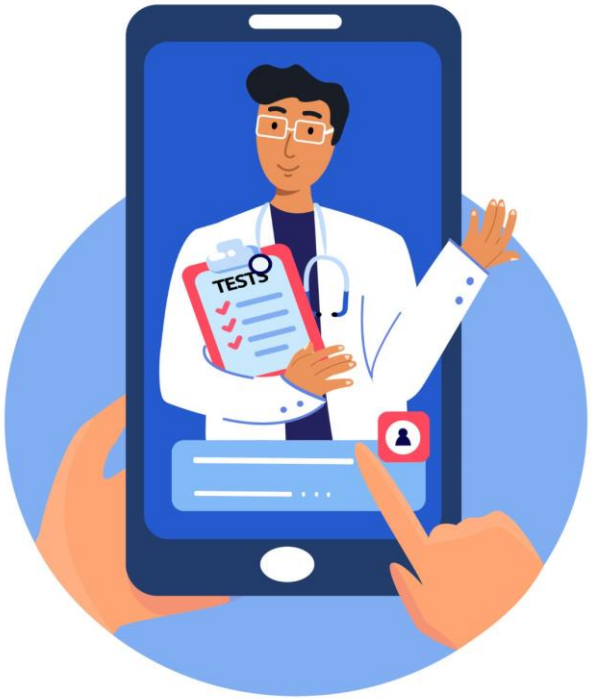
Method: rule-based text matching; large language model

- WET script detection: F1 .984 for text matching vs. .840 for an LLM
  - No false positives; runs locally with no data sent out
- Risk detection in CPT worksheets with an AI safety tool
  - Self-harm F1 .96; harm to others F1 .68 (synthetic data)
- Choose the simplest method that fits the task

AI scribes as a data source

Method: speech recognition + LLM (AI)

- Cultural idioms can be misread as symptoms
- Wrong speaker assigned in couples or group sessions
- Diagnostic language more severe than warranted
- Clients may disclose less when recorded
- Errors become part of the permanent record

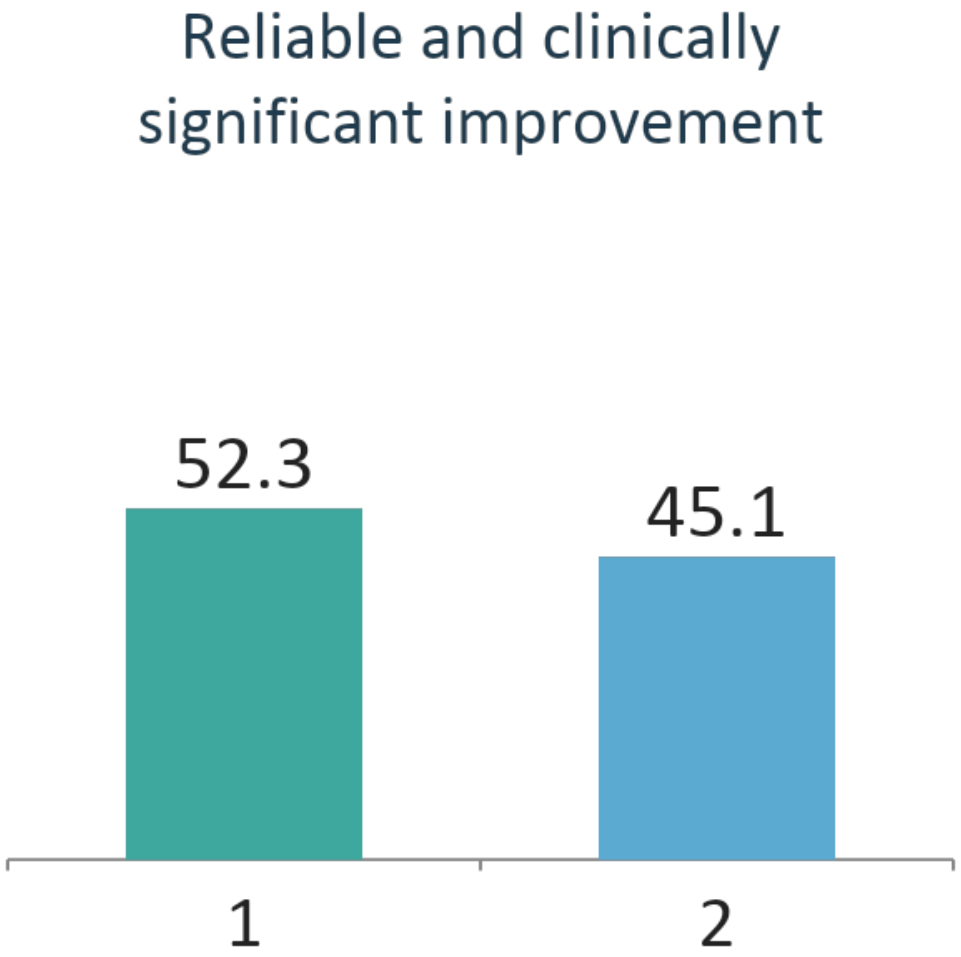


Machine learning for treatment decisions

Stratified care for depression in NHS services

Method: machine learning (AI) | Delgadillo et al., 2022

- Algorithm recommended low or high intensity at assessment
- 951 patients, cluster randomized trial
- About £104 more per patient
- Benefit in standard cases; no significant benefit in complex cases
- 95% of participants were White



Predicting who starts depression treatment

Method: ML decision tree and regression | Panaite et al., 2024

- 127,423 veterans with a new depression diagnosis
- AUC 0.59 for the regression model
- Decision tree: 88% accurate, weak at finding non-starters (F = 0.46)
- High overall accuracy can hide poor performance for key groups

Selecting Treatments for PTSD

Method: statistical, with ML variable selection | Wiltsey Stirman et al., 2021

- 267 women veterans and service members (VA CSP #494)
- Prolonged Exposure (PE) produced better outcomes for the 64% with the best prognoses
- No difference between PE and Present Centered Therapy for others
- Predictors: PTSD severity, military sexual trauma, treatment credibility, functioning

Predicting completion of PE or CPT

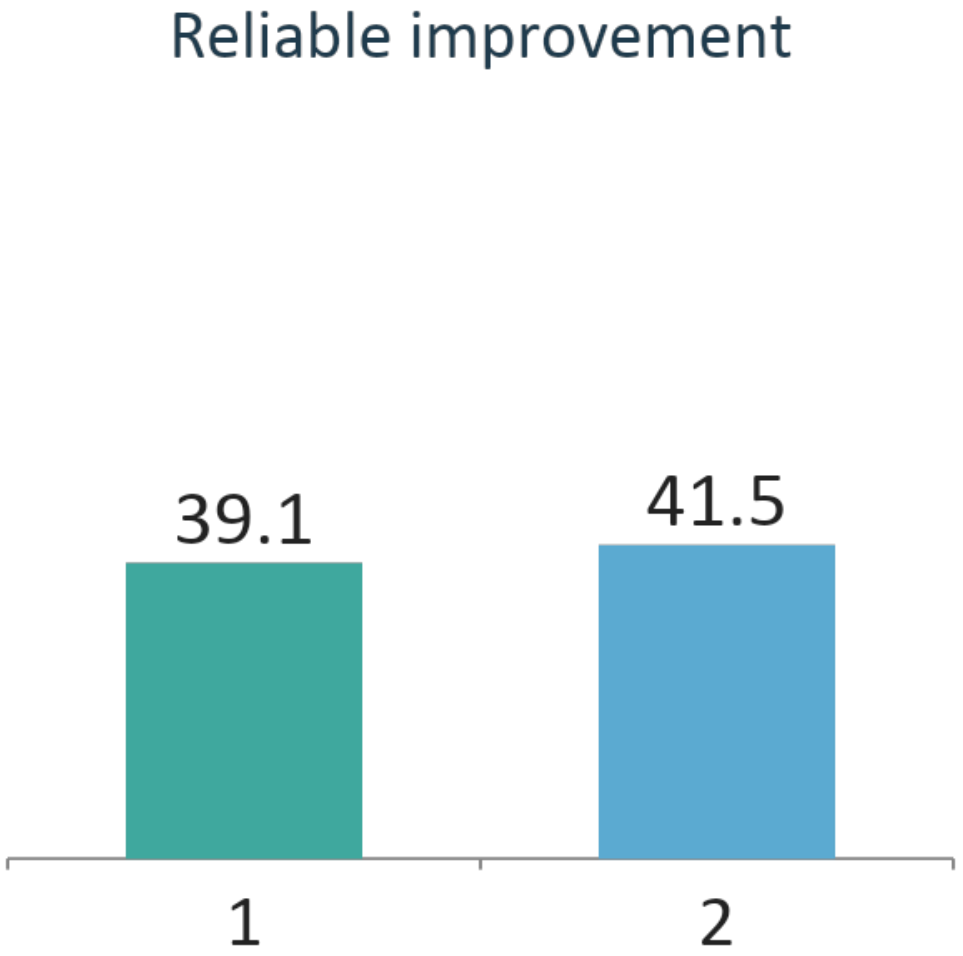
Method: statistical prediction | Keefe et al., 2018

- 160 women with rape-related PTSD
- Patients who received their model-indicated treatment had half the dropout risk (RR 0.49)
- Higher dropout risk in PE than CPT with childhood physical abuse, relationship conflict, anger, racial minority status

External test of a PTSD treatment selection index

Method: statistical prediction | Tait et al., 2024

- Index for TF-CBT vs. EMDR tested in NHS routine care
- 152 matched patients, 7 services
- No outcome difference for optimal vs. suboptimal treatment (p = .89)
- Models need testing in new settings before use



Prediction and monitoring: lessons from practice

Outcome feedback for patients not on track

Method: statistical prediction (not AI) | Delgadillo et al., 2018

- 2,233 patients in 8 NHS Trusts
- Therapists received alerts when patients were not on track
- Lower symptoms for not-on-track patients: PHQ-9 d = 0.23, GAD-7 d = 0.19
- Monitoring helps when clinicians act on the information

REACH VET: suicide risk prediction linked to outreach

Method: regression model (VA calls it AI) | McCarthy et al., 2021

- Flags the top 0.1% of predicted risk for outreach
- Evaluation in 173,313 veterans
- More safety plans and outpatient visits
- 5% fewer documented suicide attempts
- No change in suicide mortality



Suicide risk models and health system readiness

Viewpoint | Xiao, Walsh, Kharrazi, & Mann, 2026

- Pairing a real-time EHR model with the C-SSRS outperformed either alone
- No model has predicted suicide deaths across sites
- NLP of notes can find over 10 times more nonfatal suicidal behavior than diagnostic codes
- Prediction spread in sepsis and readmission through shared infrastructure: common data, EHR integration, workflow fit

Three implementation challenges

Challenge	What it requires	Proposed approach
Generalize	Stable performance over time and at new sites, including safety-net and rural clinics	Shared benchmarking of validated models across health systems
Adopt	Fit to local population, workflow, and capacity to respond	Site profile to deploy, recalibrate, reject, or choose another model
Advance	Evidence that notes, social determinants, and trajectories add value	Test before investing

Health records reflect gaps in care

Method: descriptive analysis (not AI)

- 22.8% of veterans with PTSD started an EBP; 9.1% completed (Maguen et al., 2019)
- 11.6% started an EBP in their first year; 32.9% of those received an adequate dose (Cameron et al., 2023)
- Lower odds of an adequate dose for Black, Hispanic/Latino, and Pacific Islander veterans
- Models trained on these records learn these patterns

Implementation and governance

Barriers to precision methods in routine care

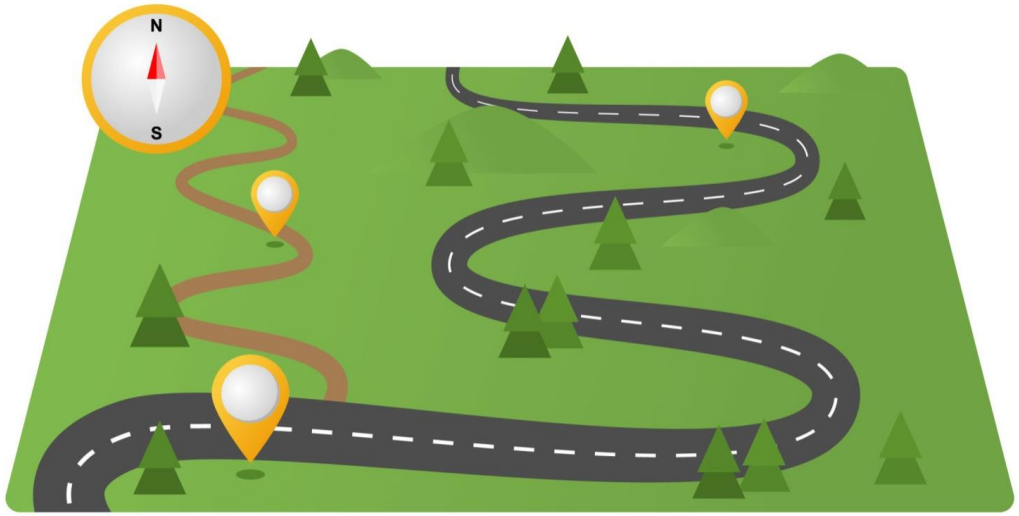
- Client burden and privacy concerns
- Clinician skepticism and limited training
- Infrastructure and cost
- Model performance drifts over time
- Models developed in narrow samples
- Biased data can widen inequities

Deisenhofer et al., 2024



Ways forward

- Co-design tools with clinicians and clients
- Start with routine progress monitoring
- Use inclusive data and check fairness
- Monitor and recalibrate models
- Test tools within routine care



Deisenhofer et al., 2024

Equity questions for any predictive tool



- Who is represented in the training data?
- How are demographic variables used?
- Are results reported by group?
- A widely used health algorithm used cost as a proxy for need and underestimated Black patients' needs (Obermeyer et al., 2019)

Consent and transparency for clinical data

- Explain what is recorded, who can access it, and how long it is kept
- Disclose whether data are used to train models
- A clinician reviews every AI-generated note
- Clients can opt out with no change in care
- Illinois requires written consent for AI recording or transcription

FDA activity, fall 2026

- No FDA-authorized generative AI mental health device
- Advisory committee (Nov 2025): human oversight, transparent labeling, postmarket monitoring
- TEMPO pilot: selected behavioral health devices offered in Medicare care without premarket authorization
- TEMPO participants added on a rolling basis

State laws and forthcoming guidance

- Some state laws govern AI interaction with clients
- Illinois prohibits AI therapeutic communication and requires written consent for AI recording
- FDA generative AI guidance forthcoming (no timeline)

Questions before adopting a predictive or AI tool

- Where was it validated, and for whom?
- How will performance be monitored over time?
- Who reviews the outputs?
- How is it disclosed to clients?
- What happens when it is wrong?



Discussion

- Which uses of data in your setting would benefit most from AI?
- What evidence would you need before acting on a prediction?
- How should clients be involved in these decisions?

References (1 of 3)

Atkins et al. (2014). Implementation Science, 9, 49

Cameron et al. (2023). Adm Policy Ment Health, 50, 813-823

Deisenhofer et al. (2024). Behav Res Ther, 172, 104443

Delgadillo et al. (2018). Lancet Psychiatry, 5, 564-572

Delgadillo et al. (2022). JAMA Psychiatry, 79, 101-108

Keefe et al. (2018). Depression and Anxiety, 35, 330-338

References (2 of 3)

Levis et al. (2026). J Clin Psychiatry, 87(2), 25m15904

Maguen et al. (2019). Adm Policy Ment Health

Maguen et al. (2019). Psychiatry Research

Maguen et al. (2021). Psychological Medicine

McCarthy et al. (2021). JAMA Netw Open, 4, e2129900

Obermeyer et al. (2019). Science, 366, 447-453

References (3 of 3)

Panaite et al. (2024). BMC Health Serv Res, 24, 529

Stade et al. (2025). Technology, Mind, and Behavior (READI)

Tait et al. (2024). Psychotherapy Research

Wiltsey Stirman et al. (2021). Behav Res Ther, 142, 103872

Xiao et al. (2026). JAMA Psychiatry, 83, 1001-1002