NATIONAL DIALOGUES ON BEHAVIORAL HEALTH, NEW ORLEANS, NOVEMBER 10, 2026

Pragmatic AI in Behavioral Health Care

Measurement, Prediction, Causal Inference, and Clinical Decision-making



Chris J. Kennedy, PhD

Center for Precision Psychiatry, Massachusetts General Hospital and Harvard Medical School

HARVARD MEDICAL SCHOOL

MASSACHUSETTS GENERAL HOSPITAL

Quick poll

> *“Social workers do not always bother to ask themselves whether the statements they make can be affirmed with certainty.”*

When was this written?

A. 2024, in guidance on using AI chatbots

B. 1985, in a clinical assessment textbook

C. 1917, in a founding casework textbook

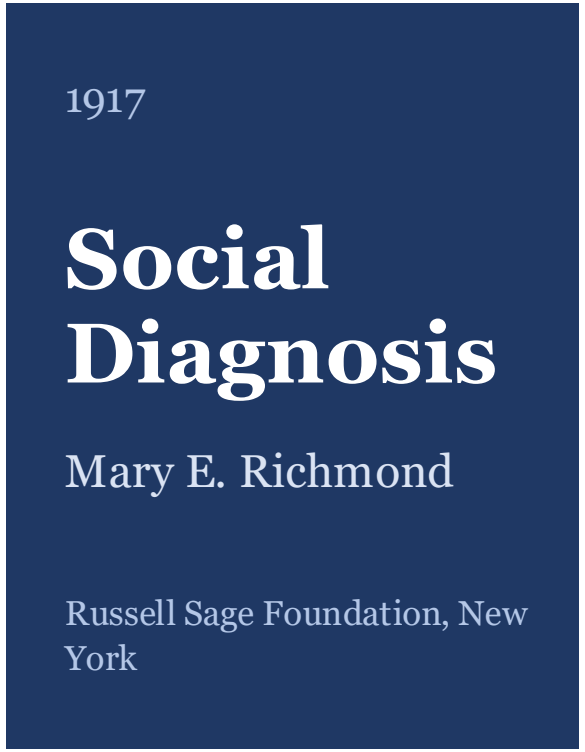
Join the polls

menti.com

1997 3032

Vote at menti.com with code **1997 3032**, or by show of hands

Social workers weighed evidence long before AI



What Mary Richmond asked caseworkers to do

- Gather evidence from the client, family, and other sources
- Ask whether each claim comes from observation or rumor
- Separate facts from inferences, and compare sources

Weighing evidence, including AI output, is a founding social work skill

Richmond, Social Diagnosis, Russell Sage Foundation, 1917

The most surprising claims need the most checking

False news spreads farther

False stories on Twitter spread farther and faster than true ones

Nonreplicable studies are cited more

Only 12% of later citations mention the failure to replicate

Checking takes seconds

Check claims during this talk with an AI search tool, then open the source

Make checking a habit, especially for surprising claims

Richmond, Social Diagnosis, 1917; Vosoughi et al., Science 2018; Serra-Garcia and Gneezy, Sci Adv 2021



Outline

1. **Surround sound measurement**
   Measurement-based care, a live demo, ambient sensing, and computer vision
2. **Causal inference**
   Doing more with the same resources, precision treatment rules, and causal diagrams
3. **Clinical decision-making**
   Suicide risk assessment, child welfare, and inpatient nursing
4. **What organizations can do**
   Ground rules, privacy, linked data, AI agents, and evaluation

In 1997, a computer beat the world chess champion

1997 IBM's Deep Blue defeats Garry Kasparov

1998 Kasparov starts matches where each player has a computer

2005 Two amateurs using three computers beat grandmasters in an open tournament

What made the difference?



Kasparov, New York Review of Books 2010

Kasparov's Law

Weak human + machine
+ strong process

beats

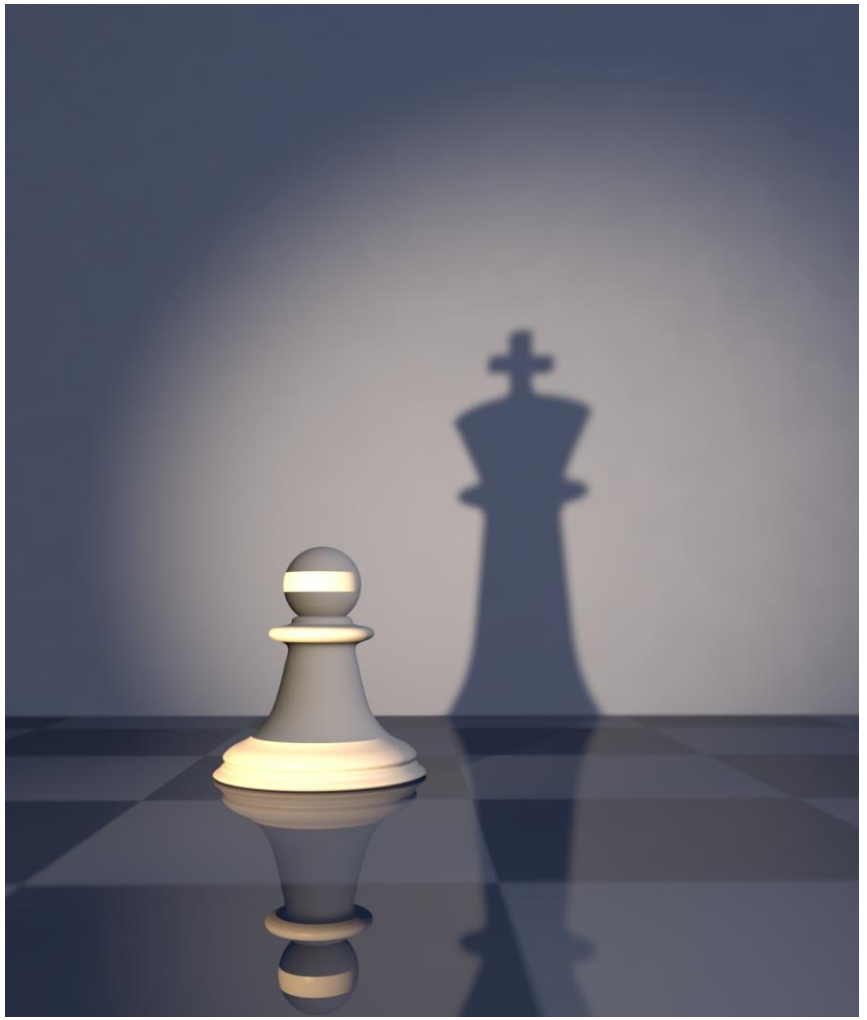
Strong human + machine
+ weak process

Marwaha and colleagues, summarizing Garry Kasparov

OUR GOAL, THROUGH EFFECTIVE TRAINING

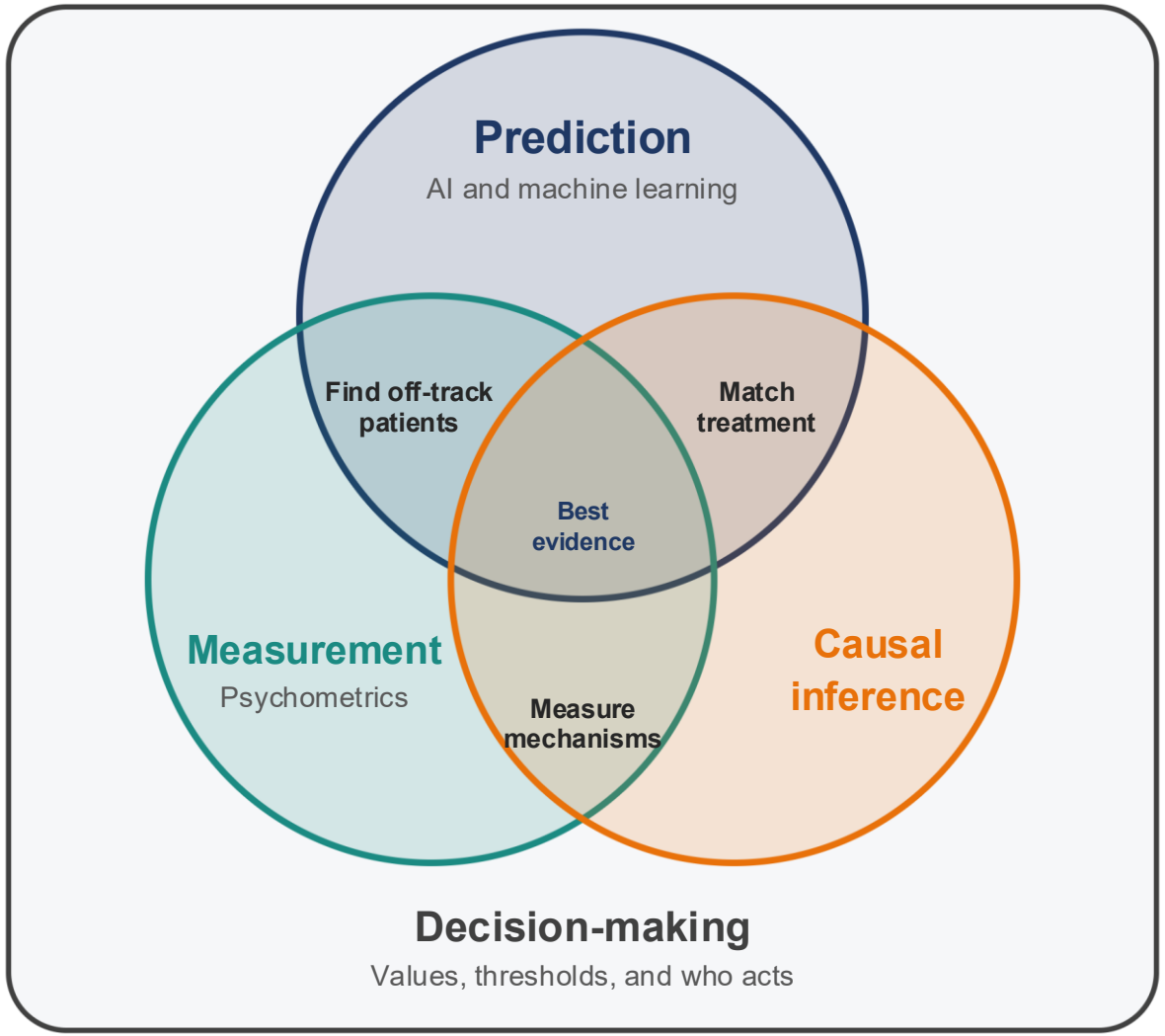
Strong human + machine + strong process

A strong process can mean clinicians record their own assessment before seeing the model's estimate, then review every case where the two differ



Marwaha et al., Machine Learning for Health (ML4H) 2025; Kasparov, New York Review of Books 2010

AI is one of three tools for better decisions



- AI predicts what is likely to happen
- Choosing a treatment requires knowing what would happen under each option
- Every model depends on valid measures
- Decisions add values, thresholds, and who acts

Ask how any AI tool handles all three, and who acts on its output

PART 1

Surround sound measurement

Even weekly therapy covers less than 1% of a patient’s month

New evidence from the hours between visits

Our current measurement projects

CONVERSATION

Depression monitoring (AURORA trial)

Severity between visits from conversational check-ins with an AI agent

IMAGES

Self-injury severity

Computer vision rates self-injury scars in photos

CLINICAL NOTES

Sentiment and social needs

Natural language processing reads clinical notes

NEIGHBORHOODS

Area-level context

Addresses linked to neighborhood measures

RATING SCALES

Psychiatric acuity

Item response models separate patient severity from rater habits

SMALLER GROUPS

Fairness

Transfer learning improves accuracy for groups with little data

Center for Precision Psychiatry, Massachusetts General Hospital, with collaborators

The AURORA trial tests companion AI between visits

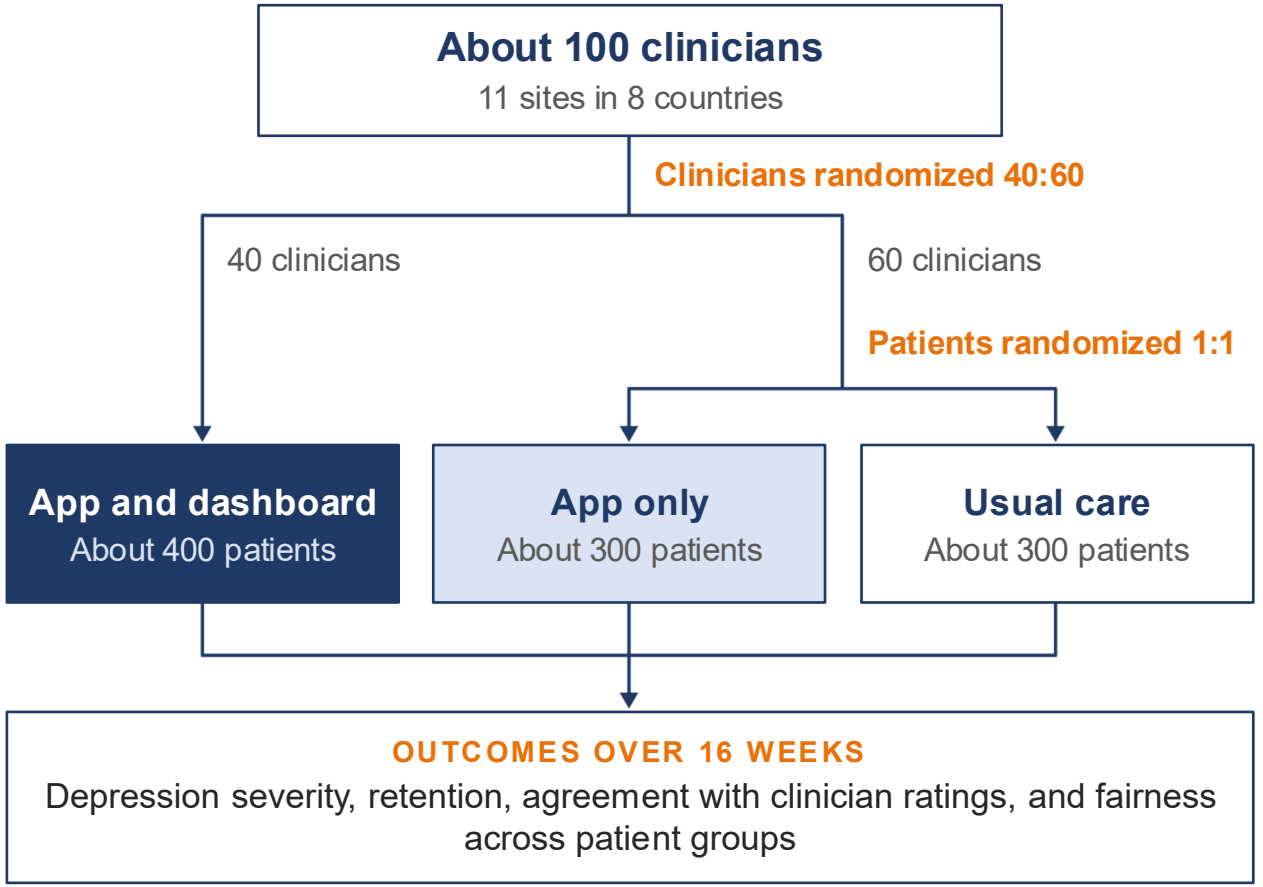
HOW IT WORKS

Patients in the two app arms complete conversational check-ins with an AI agent every 1 to 3 days, and their phones collect passive data.

Models estimate depression severity from the check-ins.

Clinicians in the dashboard arm see severity trends and alerts.

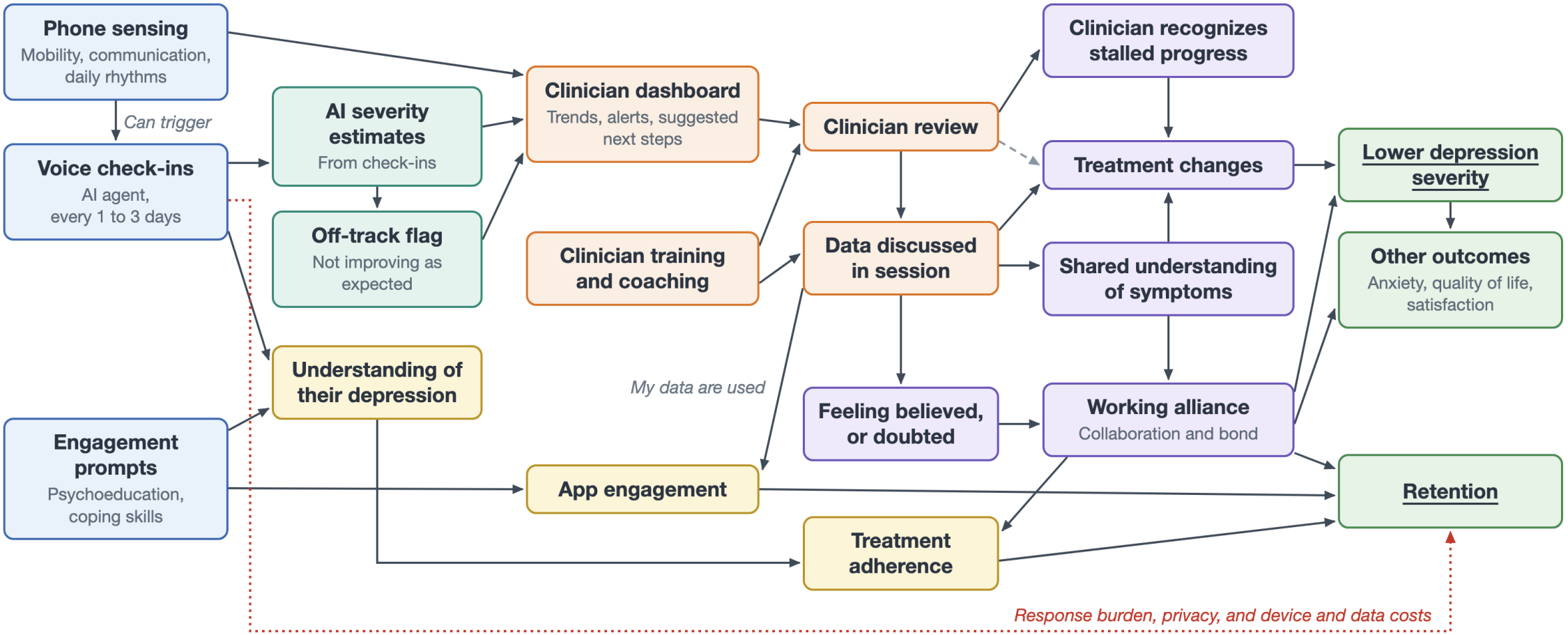
TRIAL DESIGN



Funded by Wellcome (332864/Z/25/Z); contact principal investigator Eric Storch, Baylor College of Medicine; July 2026 to June 2028

The AURORA trial's conceptual model

Conceptual models provide structure to empower intervention understanding and effectiveness



Live demo: AURORA clinician dashboard

Demonstration data only

What to look for

- Each patient’s estimated severity between visits
- Alerts when symptoms worsen
- Which of these alerts would change what you do this week?

Open the live demo

From measuring between visits to helping between visits

A triggered nudge is support that arrives when a check-in, a location, or a sensor suggests someone may need it



An example from addiction care

Risky drinking days in the past 30 days



Trial of 349 people leaving residential treatment for alcohol use disorder

One counselor spent about 2 hours a week following 120 patients

Micro-randomized trials test which nudges help, and when

The prompt wording is illustrative. Gustafson et al., JAMA Psychiatry 2014; Klasnja et al., Ann Behav Med 2019; Nahum-Shani et al., Ann Behav Med 2018

Measuring sleep and breathing with radio waves

A device in the room measures radio reflections, with no camera or wearable



What researchers have measured

80.5% agreement with laboratory sleep studies on sleep stages, in 880 people

11 of 11 seizures showed signal changes at onset, in an early study

Parkinson's disease
detected from breathing during sleep

Staff could check patients overnight without waking them

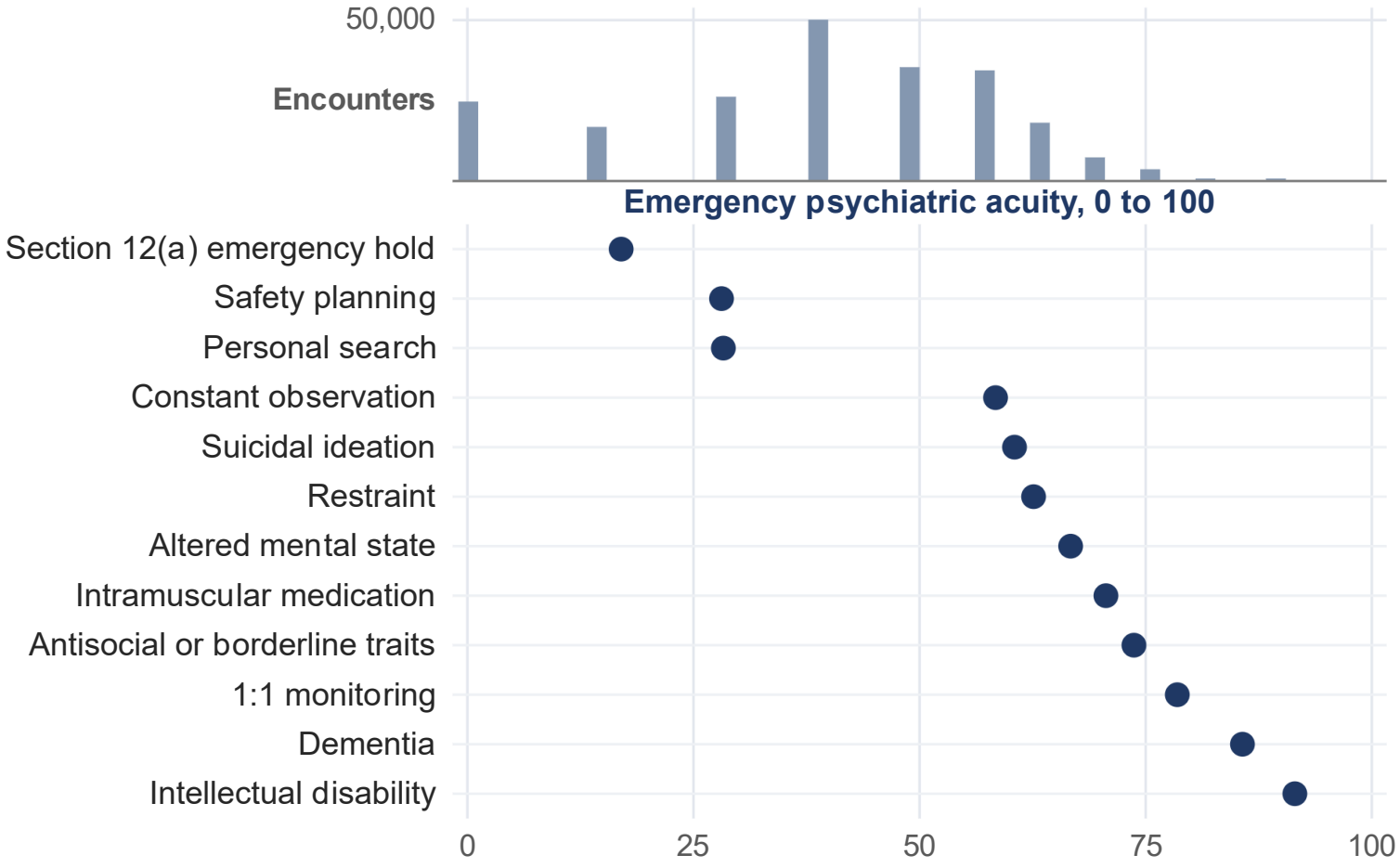
Staff could track sleep every night, since poor sleep raises suicide risk

It still needs testing on psychiatric units, with patient consent

He et al., Sleep 2025; Yang et al., Nat Med 2022; Lemus et al., Epilepsy Behav 2026; Bernert et al., Curr Psychiatry Rep 2015

Item response theory with emergency department records

Health records can measure psychiatric acuity



- Safety planning becomes likely even at low acuity
- 1:1 monitoring becomes likely only at high acuity
- Most encounters have moderate acuity

Staff already record these events, so the scale adds no new forms

Adapted from Figure 2 of our item response theory (IRT) acuity project. Section 12(a) is a Massachusetts application for a 3-day involuntary hospitalization.

Self-injury and suicide risk

Self-injury scars are linked to suicide risk

Among 78 college students who had self-injured

Had at least one lasting scar



55%

Had ever received medical care for self-injury



9%

Most people with scars had never received medical care for self-injury

231 college students

Having scars, and having more of them, predicted current suicidal thoughts beyond depression, past attempts, and how often and how severely people self-injured

115 adolescent patients, standardized photos

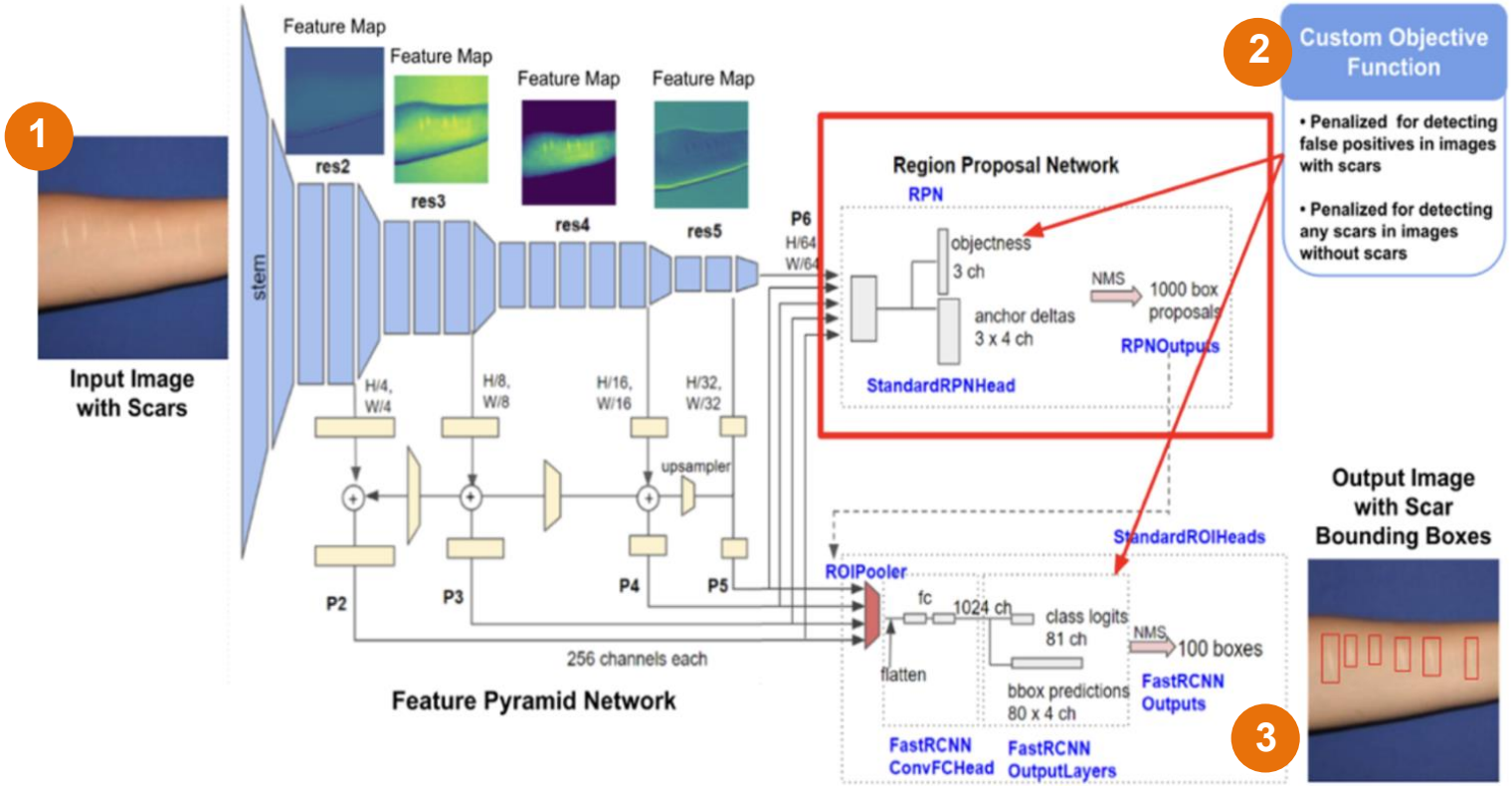
A scar score was linked to suicide risk, but on its own it predicted risk poorly

Measuring scars consistently could add to suicide risk assessment

Burke et al., Compr Psychiatry 2016; Shi et al., BMC Psychiatry 2026

Computer vision for self-injury scars

We are training a model to count scars in photos



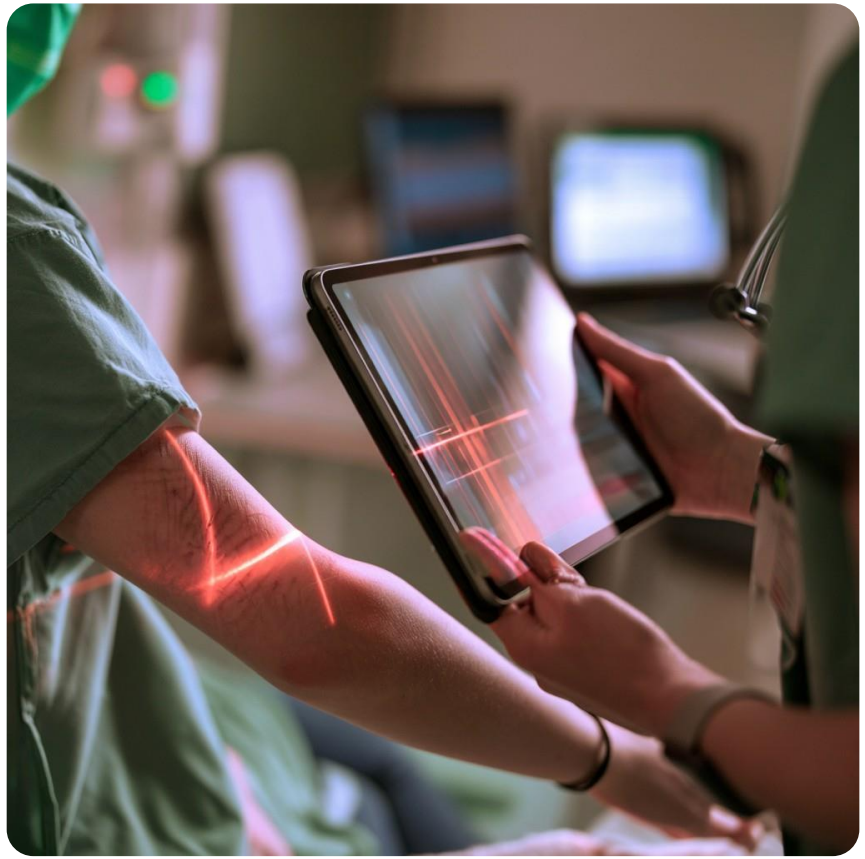
1. A photo of healed scars, taken with consent
2. Training penalizes boxes drawn where there is no scar
3. The model draws a box around each scar and counts them

A clinician reviews the boxes before the count informs a risk estimate

Figure: Faster region-based convolutional neural network (Faster R-CNN) with a custom training objective, from our project; Daneshjou et al., Sci Adv 2022

Computer vision for self-injury scars

We are piloting scar photos in emergency departments



- We are collecting images of self-injury scars during emergency department visits
- With computer vision, we can estimate scar characteristics that reflect the severity of suicidality, such as recency and method
- The goal is suicide risk prediction that does not depend on self-report alone

Illustration, not a photo from the pilot



PART 2

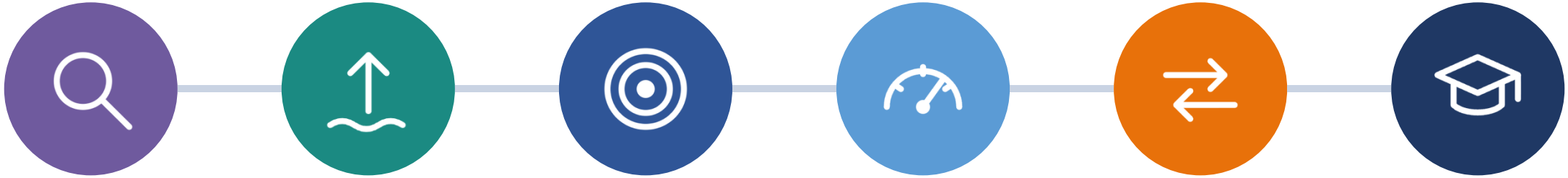
Causal inference

Knowing what would happen under each option, for each client

Doing more with the same resources

The question at this conference

How can we do more with the same resources?



Identify risk

Move upstream

Match interventions

Measure mechanisms

Switch early

Train

These steps put measurement-based care into practice and improve the outcomes that value-based care pays for

Moving upstream in REWRITE, our posttraumatic stress disorder (PTSD) prevention program

Preventing PTSD before avoidance becomes entrenched

1

Identify

A health record model identifies emergency department trauma survivors at highest risk

2

Intervene early

Offer a written exposure intervention soon after the trauma

3

Measure mechanisms

Track avoidance to check that the intervention is working

4

Extend

Bring the approach to bereavement, sexual assault, and primary care services

REWRITE combines four of the six steps: identify risk, move upstream, match interventions, and measure mechanisms

Our PTSD risk model, built on health record data

The model uses many kinds of record data

Clinical history
312 features: labs, medications, procedures

Demographics and insurance
9 features: age, sex, employment, payer

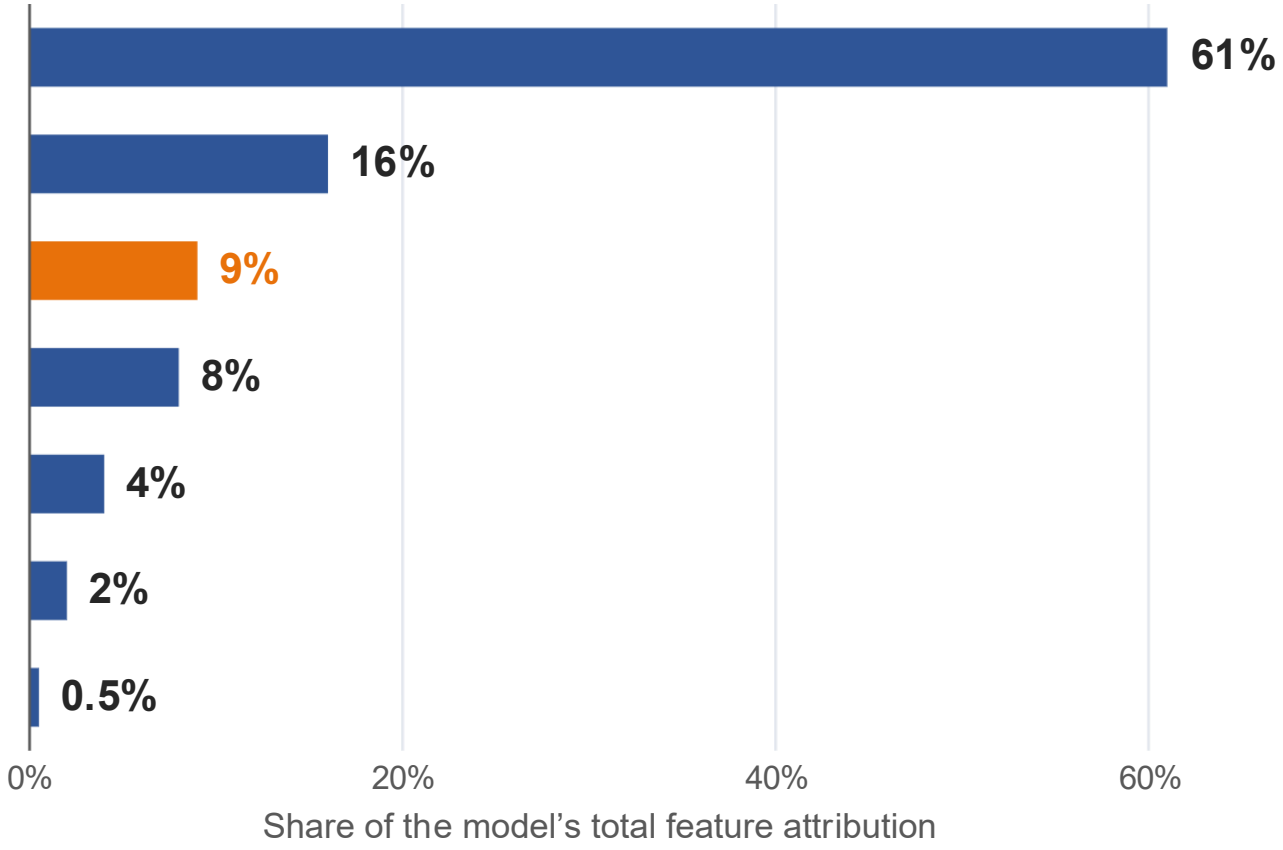
Neighborhood
38 features: income, poverty, deprivation

Injury visit
25 features: injury severity, violence screen

Zatzick health record screen
3 features: an existing 0 to 9 score

Visit history
6 features: prior visits and stays

Depression screeners
3 features: Patient Health Questionnaire

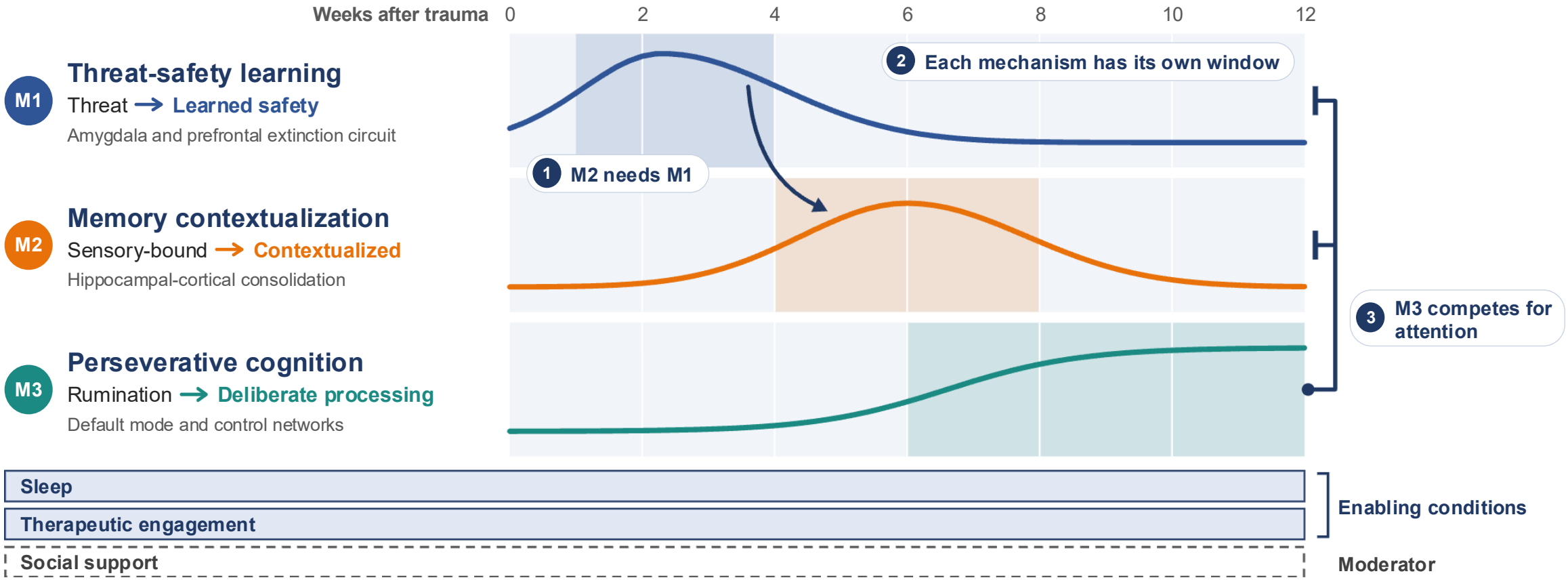


Neighborhood data contributed about as much as the injury visit itself

Adapted from Figure 4 of our PTSD risk model manuscript; Russo et al., Gen Hosp Psychiatry 2013

The theory behind REWRITE, Mechanism-Aligned Intervention Theory (MAIT)

Recovery requires change in three mechanisms



Adapted from Figure 1 of the MAIT manuscript, in preparation. Curves show predicted activation of each mechanism's PTSD-maintaining pole.

Precision treatment rules

Match at intake, then measure whether treatment works

Step 1

Match

A rule uses each client's intake data to recommend the treatment most likely to help

Step 2

Measure

Track the mechanism the treatment targets, such as avoidance, between sessions

Step 3

Switch early

When the mechanism is not changing, reassign treatment within weeks

Each client's results improve the rule

In REWRITE, we predict that each person's mechanism profile shows whether written exposure therapy or Brief STAIR will help more

Causal inference in routine care

Routine outcome data can answer three causal questions

How effective is each therapist?

Adjust for each therapist's mix of clients

EFFECT SIZE

0.75

Clients matched to therapists with strong records for their problems improved more

Which treatment will work best for this client?

Compare each client's likely outcome under each treatment

EFFECT SIZE

0.58

Patients randomized to the treatment a model favored did better

Is this client off track?

Compare progress with the expected path

EFFECT SIZE

0.23

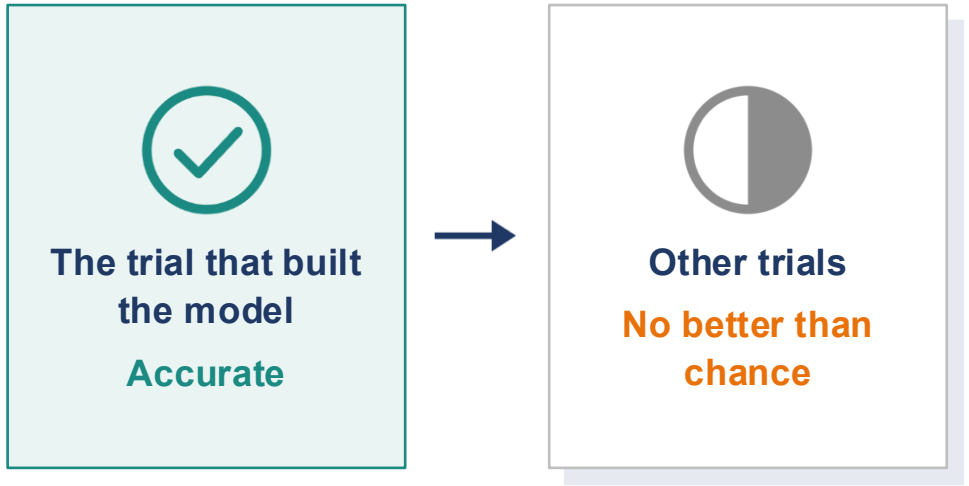
Off-track patients whose therapists got alerts ended with milder depression

All three answers depend on accounting for who received what, and why

Constantino et al., JAMA Psychiatry 2021; DeRubeis et al., PLoS One 2014; Delgadillo et al., Lancet Psychiatry 2018

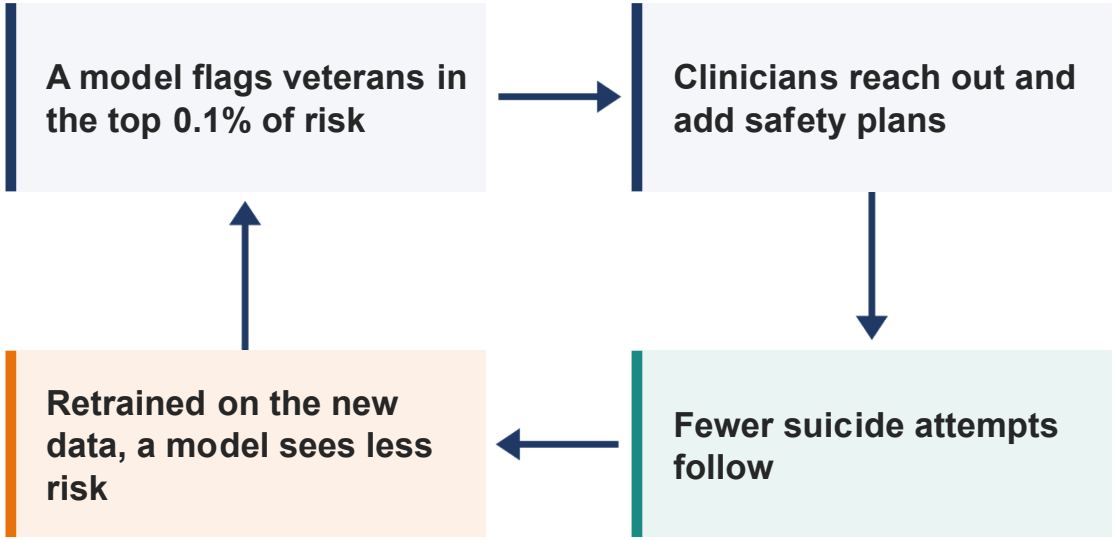
Mechanisms keep predictions accurate in new settings

Patterns from one setting can fail in another



Models predicting antipsychotic response in schizophrenia trials

Acting on predictions changes the data



Suicide prevention outreach in the Veterans Health Administration

Before trusting a predictor, ask what process links it to the outcome

Chekroud et al., Science 2024; McCarthy et al., JAMA Netw Open 2021; Sperrin et al. 2019; Dickerman and Hernán 2020; Subbaswamy and Saria 2020

Causal diagrams

A model trained on past care underestimates risk for patients who were treated



A model trained on past care

Learns risk after treatment, so high-risk patients who were treated look safer

Counterfactual prediction

Estimates risk under each option, such as continuing, adjusting, or escalating care

Drawing the diagram shows what to measure to predict risk under each option

Clinicians need routine skill monitoring too

Step 1

Rate skills

Rate each clinician's skills in recorded or simulated sessions

Step 2

Compare

Compare progress with an expected learning curve

Step 3

Train

When someone falls off track, add AI feedback and simulated practice

Across 170 therapists, client outcomes declined slightly with experience

Practice time predicted client outcomes, and experience did not

Supervisors can give every clinician the feedback loop that top performers build for themselves

Goldberg et al., J Couns Psychol 2016; Chow et al., Psychotherapy 2015

Small-group challenge, 4 minutes

Smaller budget, same clients

Next year, your agency's budget is cut by 20% and referrals stay the same. With the 2 or 3 people next to you, pick one of these steps and make a plan.


Identify risk


Move upstream


Match interventions


Measure mechanisms


Switch early


Train

1. What would you try first?
2. What could go wrong for clients?
3. How would you know it worked?

Then one person from each group shares one sentence



PART 3

Clinical decision-making

The clinician assesses risk, a model combines that rating with health record data, and the clinician decides

Weighing a model’s estimate alongside the clinician’s judgment

Quick poll

Earlier, I said our attention spans have dropped to 8 seconds, shorter than a goldfish's.



What did you do?

A. I believed it

B. I doubted it, but did not check

C. I checked it

Vote at menti.com with code **1997 3032**, or by show of hands

No study supports the goldfish statistic



1. A 2015 Microsoft Canada report claimed 8 seconds, below a goldfish
2. The report credits a statistics website, not Microsoft research
3. A BBC reporter found no research behind it

> *“Social workers do not always bother to ask themselves whether the statements they make can be affirmed with certainty.”*
>
> Mary Richmond, 1917

Maybin, BBC World Service 2017; Microsoft Canada 2015; Bradbury, Adv Physiol Educ 2016

Clinical vs. statistical prediction

Clinicians

- Draw on the interview, the patient's own account, and context the record never captures
- Weight the same information differently from patient to patient and from clinician to clinician

Statistical models

- Apply the same weights to every patient, every time
- Use only what the record contains

- Since Meehl's 1954 book, researchers have usually found statistical prediction as accurate as clinical prediction or better
- Child welfare researchers have had the same debate about actuarial vs. consensus-based risk assessment
- Recent work asks how to combine the two, and in what order

Meehl 1954; Dawes et al. 1989; Grove et al. 2000; Baird and Wagner 2000; Simon et al. 2021; Roske and Nock 2025

In an assess-first process, the clinician rates risk first

1 Clinician assesses

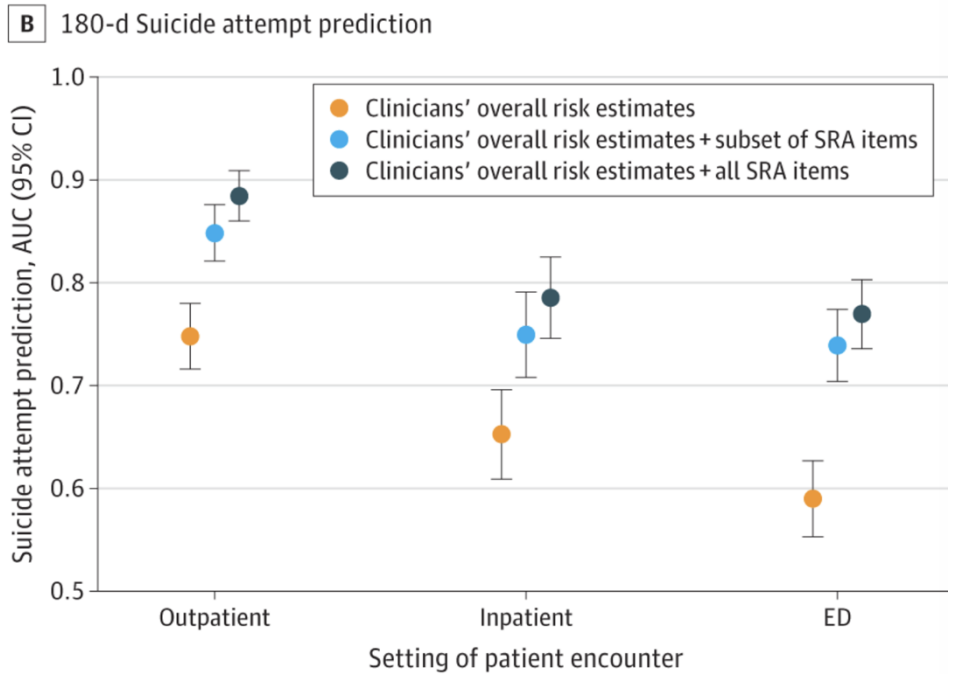
Rates risk before seeing the model, including what the record cannot show

2 Model combines

Combines the assessment items, the rating, and health record data

3 Clinician decides

Checks for errors, reviews disagreements, and sets the plan





Adding assessment items to the rating improved prediction in every setting

Bentley et al., JAMA Psychiatry 2025. AUC, area under the curve; CI, confidence interval; SRA, suicide risk assessment; ED, emergency department; 180-d, 180 days

Why the clinician assesses first

Independent judgment

The clinician rates risk before seeing any model output, so the model cannot influence the rating

Expertise as input

Clinicians can record one or more independent ratings, and the model uses each one

Better use of the estimate

Showing how the clinician’s ratings shaped the estimate can improve how clinicians use it

Learning from disagreement

With both estimates on record, teams can review cases where they differ

The clinician’s rating was among the most important predictors in every setting

Clinicians gain from a prediction only when it changes their plan

	Model estimates low risk	Model estimates high risk
Clinician plans routine care	AGREEMENT The plan is unchanged	DISAGREEMENT **Consider escalating, which helps only if risk is truly high**
Clinician plans intensive care	DISAGREEMENT **Consider a less intensive plan, which helps only if risk is truly low**	AGREEMENT The plan is unchanged

Train clinicians to review these disagreements and decide how much weight to give the model's estimate

Marwaha et al., Machine Learning for Health (ML4H) 2025

Case vote

You completed a suicide risk assessment, and a model then estimated this patient's risk

Your rating: **Minimal**

Model estimate: **10.3% chance of a suicide attempt within 180 days**

Average across all assessments: **0.8%**

What do you do?

A. Keep the current plan

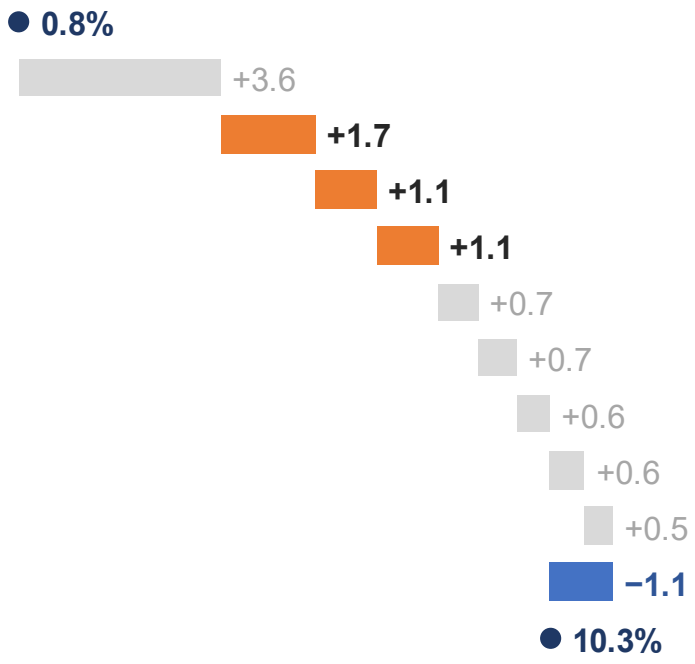
B. Reassess the patient, then adjust the plan

C. Escalate care now

Talk it over with a neighbor for 20 seconds, then vote at menti.com with code **1997 3032**, or by show of hands. Illustrative case from our suicide risk study funded by the National Institute of Mental Health

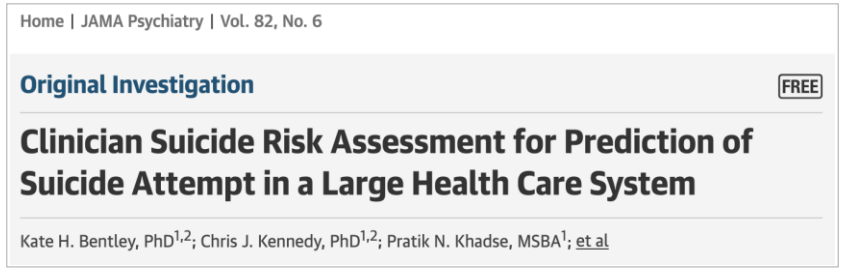
Why the model's estimate differs

Each factor's contribution to this patient's predicted 180-day risk (percentage points)



Summary drafted by a language model

Average suicide attempt risk is 0.8% but the model predicts this patient's individual risk as 10.3% within 180 days after a clinical suicide risk assessment. Key factors increasing risk include substance use/intoxication (+1.7 percentage points), sense of responsibility to family (+1.1%), chronic medical illness (+1.1%), and depressed mood (+0.7%). The only factor that reduced risk was the clinician's overall risk estimate of 1 ["minimal"], which decreased the predicted risk by -1.1%.


Home | JAMA Psychiatry | Vol. 82, No. 6

Original Investigation FREE

Clinician Suicide Risk Assessment for Prediction of Suicide Attempt in a Large Health Care System

Kate H. Bentley, PhD[1,2]; Chris J. Kennedy, PhD[1,2]; Pratik N. Khadse, MSBA[1]; et al

Bentley et al., JAMA Psychiatry 2025. The patient example is illustrative.

Building skill through deliberate practice

- Rating categories with explicit behavioral and probability anchors, as in the Veterans Affairs risk stratification table
- Outcome feedback on past or simulated cases, so clinicians can compare their ratings with what happened
- Review of cases where the clinician’s rating and the model’s estimate differ, with the reasoning documented
- When clinicians rate more accurately, the model’s estimates improve too, since their ratings are among its top inputs

Practicing hard situations in simulation

Virtual reality therapy for clients

- Clients complete about 6 sessions of automated virtual reality therapy (gameChange), with a staff member present
- In a trial of 346 people with psychosis, participants reported less avoidance and distress in everyday situations
- England's National Institute for Health and Care Excellence recommends it for NHS use while more evidence is gathered

AI-simulated patients for clinicians

- Trainees interview a patient simulated by a language model and get structured feedback
- Psychiatry residents found the simulations plausible, but rapport with the simulated patient was hard to build
- Social workers and nurses could use the same approach to rehearse risk assessment and de-escalation

The largest gains were among people with the most severe avoidance, such as those who could not leave home alone

Freeman et al., Lancet Psychiatry 2022; Greater Manchester Mental Health NHS Foundation Trust 2023; Rueda et al., JMIR Med Educ 2026

Social workers already make decisions with algorithms

About 537,000 social workers form the largest US mental health profession

Child welfare hotline screening

Allegheny County, Pennsylvania

Screeners see a risk score when they decide whether a hotline report is investigated

Gap between Black and White children screened in for investigation



Homelessness prevention

Los Angeles County

- Staff use a model to contact people at risk of losing housing early
- A randomized trial is due to finish in 2027

Screeners who weighed the score against what else they knew cut the gap between Black and White children by more than half

HRSA 2024 data via APNA 2025; Cheng et al., CHI 2022; California Policy Lab 2025; Human Rights Data Analysis Group 2023

Psychiatric nurses and AI on inpatient units

About 160,000 psychiatric nurses form the second-largest US mental health profession

⚠ Predicting violence on the unit

- Researchers predicted violence in the first 4 weeks from first-day notes
- Accuracy dropped when researchers used a model at the other hospital

Guidance from nursing

- Nurses remain accountable for their decisions, even when technology fails
- Nurses should help govern AI

Before switching on any alert, ask the night staff which ones they would act on

American Psychiatric Nurses Association 2025; Menger et al., JAMA Netw Open 2019; American Nurses Association position statement, OJIN 2025

PART 4

What organizations can do

Ground rules, AI agents, and learning from every rollout

Deciding which AI tools to trust, and checking them after rollout

Quick poll

In a web browser, you share files with an AI tool by uploading them. In a desktop app, such as the Claude desktop app, you can also let it open and create files in folders on your computer.

Where do you use AI tools for work most often?

A. In a web browser

B. In a desktop app that works with files on my computer

C. Inside programs I already use, such as Microsoft 365 Copilot

D. I do not use AI tools for work

Vote at menti.com with code **1997 3032**, or by show of hands

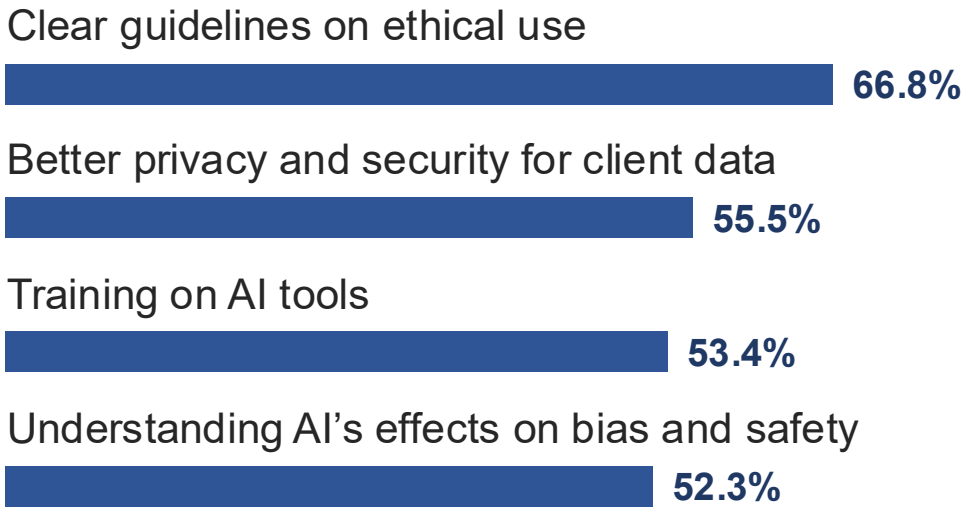
Most social workers already use AI at work

63.5%

of 1,179 social workers in a national survey use AI in their current role

- Top uses are correspondence, reports, and clinical documentation
- Top concerns are privacy, security, and ethics
- Most said their agency offers little AI guidance

What they said would help



Borah et al., Moritz Center for Societal Impact, University of Texas at Austin, 2026; University at Buffalo 2026

Ground rules for staff use of AI

1 Approved tools only

Only tools with a business associate agreement get client information

2 Tell clients

Tell clients when staff use AI, and ask for consent

3 Read every note

Clinicians read and correct every AI draft before signing

4 Train first

Train staff on where the tool fails, and review the policy yearly

Based on concerns reported by Borah et al. 2026 and the National Association of Social Workers 2026

Using language models while protecting client privacy

Commercial models under a business associate agreement

- The vendor signs an agreement to protect client data
- OpenAI, Anthropic, Google, and Microsoft offer them for enterprise products
- Data still goes to the vendor, so configure it carefully

Open-weight models on the organization's own servers

- Run a downloaded model, such as Llama or Gemma, on your own servers
- Client data stays inside the organization's network
- It requires graphics processing units (GPUs) and technical staff

Use client information only in tools approved through one of these routes

HHS guidance on HIPAA and cloud computing; vendor documentation, 2026; Buckley et al., JAMA Health Forum 2025; Wiest et al., Br J Psychiatry 2024

Linked data show outcomes that one health system's records miss

What one health system's records miss

Attempts treated elsewhere

These usually appear in insurance claims, including Medicaid, but not in the treating system's own records.

Suicide deaths

Deaths are recorded in state vital records, and they are too rare in any one system to model well.

Differences across systems

Assessment forms, documentation habits, and case mix differ across systems, so a model built in one may not transfer.

Two privacy-preserving remedies

Privacy-preserving record linkage

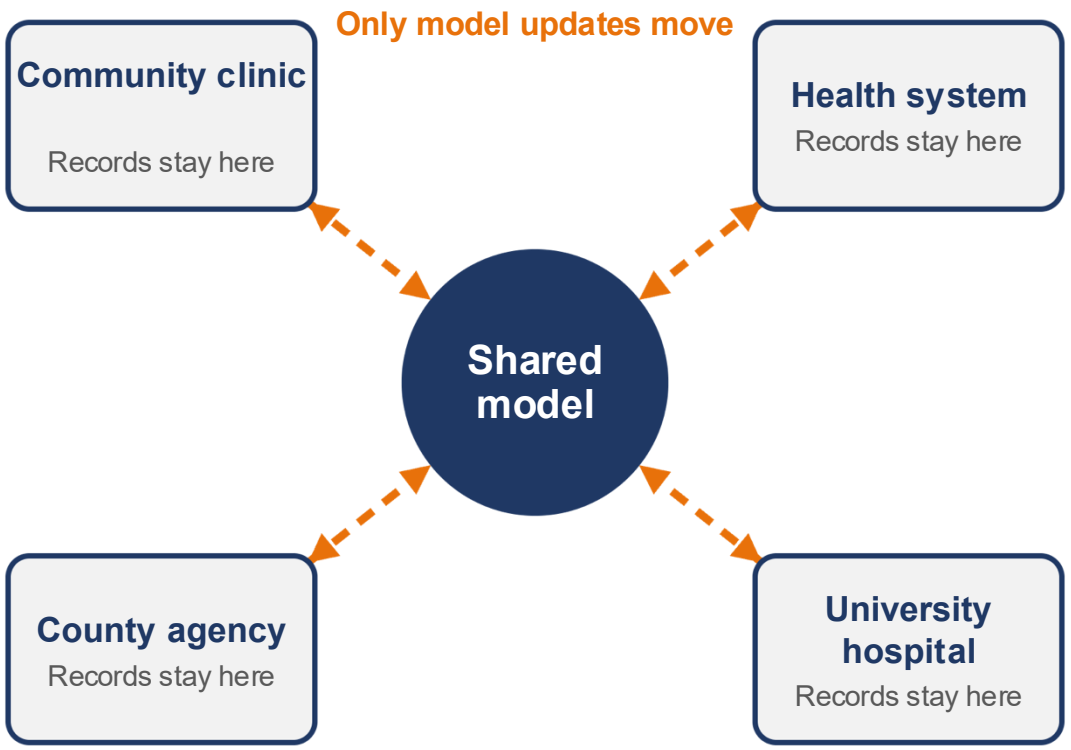
Cryptographic tokens link hospital, claims, and death records across systems without sharing names or other direct identifiers.

Federated learning

Each system trains a model on its own data and shares only the model updates, so researchers can build one model from many systems' attempts and deaths.

States can make vital records and Medicaid claims available for privacy-protected linkage

Sites can train one model without pooling records



- Larger samples, with enough people to check fairness for small groups
- Models that stay accurate across sites, 16% more accurate in a 20-institution study
- Enough data to estimate who benefits from which treatment

State and county systems can build shared models while client records stay on site

Dayan et al., Nat Med 2021

When AI agents take actions

Agentic AI completes multistep tasks, such as booking appointments or submitting forms

What agents can do now

15% vs. 6% Rise in referrals when a chatbot handled self-referral, compared with other services

179% Rise among nonbinary people, and 29% among ethnic minority groups

70% Best agent's success rate on 300 realistic health record tasks (2025)

Match autonomy to the level of risk

Drafts a person finishes

Notes, letters, and referral forms

Actions a person approves

Messages, scheduling, and prior authorizations

Never without a person

Risk assessment, crisis response, and reports to child protective services

Decide in advance which actions an agent may take without a person

Habicht et al., Nat Med 2024; Jiang et al., NEJM AI 2025

Evaluate AI tools the way we evaluate treatments

Randomize the rollout

By clinic or clinician, as in the AURORA trial

Track outcomes that matter

Suicide attempts and deaths, engagement, and access to care, alongside symptom scores

Monitor after launch

Accuracy across groups over time, and how often clinicians override it

If rollout is staggered anyway, randomizing the order costs little and shows whether the tool helps

Quick poll

Which idea are you most excited to try?

A. Moving upstream to prevention

B. Measuring between visits

C. Assess-first risk prediction

D. Matching treatment in advance

E. Federated learning across systems

F. AI-supported training

Vote at menti.com with code **1997 3032**, or by show of hands

Summary

MEASUREMENT

The AURORA trial

Clinicians see depression severity between sessions, estimated from AI-guided conversational check-ins, in 8 countries

MEASUREMENT

Measures from routine care

An emergency acuity scale built from events staff already record, and self-injury scars rated in photos

PREDICTION

Suicide risk assessment

Adding assessment items to clinicians' ratings raised the area under the curve from about 0.6 to 0.77 in emergency care

PREDICTION

PTSD risk after trauma

In our health record model, neighborhood data contributed about as much as the injury visit itself

CAUSAL INFERENCE

Precision treatment rules

In REWRITE, mechanism profiles may show whether written exposure therapy or Brief STAIR helps each person more

CLINICAL DECISIONS

Assess-first process

Clinicians rate risk before seeing the model, then review its estimate with each factor's contribution and its uncertainty

Bentley et al., JAMA Psychiatry 2025. AURORA is funded by Wellcome (332864/Z/25/Z).

What you can do next week

1. Review off-track clients in supervision, using one routine outcome measure
2. Ask compliance which AI tools have a business associate agreement
3. Have clinicians rate risk before seeing model output
4. Randomize the order of your next rollout, and name its outcome
5. Add a social worker, a nurse, and a person with lived experience to AI governance

Mary Richmond, Social Diagnosis, 1917

“The case worker’s best safeguard against formalism and skepticism is a concern for the interests of his client.”

A century later, we have new formulas and the same safeguard. Pick one of the five actions and start it next week.

Thank you

Collaborators

Kate Bentley, Kace Boland, Jasmin Brooks Stephens, Taylor Burke, Matthew Flics, Becky Fortgang, Hannah Frank, Ron Kessler, Pratik Khadse, Hyunjoon Lee, Richard Liu, Emily Madsen, Matt Nock, Chloe Roske, Jordan Smoller, Eric Storch, Qingqing Yin, and Sasha Zitter

Funding

National Institute of Mental Health (K01 MH135131, R01 MH143822, K23 MH120436, K23 MH126168, P50 MH129699); Wellcome (332864/Z/25/Z); Massachusetts General Hospital Claflin Distinguished Scholar Award; Tommy Fuss Fund

Contact

ckennedy13@mgh.harvard.edu

References

American Nurses Association. ANA position statement: the ethical use of artificial intelligence in nursing practice. OJIN: The Online Journal of Issues in Nursing. 2025;30(2). doi:10.3912/OJIN.Vol30No02PoSCol03

American Psychiatric Nurses Association. The state of the psychiatric-mental health nursing workforce. October 2025.

Anthropic. Business associate agreements (BAA) for commercial customers [help center article]. Accessed November 9, 2026.

Baird C, Wagner D. The relative validity of actuarial- and consensus-based risk assessment systems. Children and Youth Services Review. 2000;22(11-12):839–871. doi:10.1016/S0190-7409(00)00122-5

Bentley KH, Kennedy CJ, Khadse PN, et al. Clinician suicide risk assessment for prediction of suicide attempt in a large health care system. JAMA Psychiatry. 2025;82(6):599–608. doi:10.1001/jamapsychiatry.2025.0325

Bernert RA, Kim JS, Iwata NG, Perlis ML. Sleep disturbances as an evidence-based suicide risk factor. Current Psychiatry Reports. 2015;17(3):15. doi:10.1007/s11920-015-0554-4

Blackwell B, Caprara C, Rountree J, Casey P, Vanderford D, Battis C. Early outcomes from the Los Angeles County Homelessness Prevention Unit. California Policy Lab; July 10, 2025.

Borah E, Meyerhoff J, Al-Turk A, Gower K, Mastryukova A. Use of artificial intelligence in social work practice: findings and recommendations from a national survey. Moritz Center for Societal Impact, The University of Texas at Austin; June 2026.

Bradbury NA. Attention span during lectures: 8 seconds, 10 minutes, or more? Advances in Physiology Education. 2016;40(4):509–513. doi:10.1152/advan.00109.2016

Buckley TA, Crowe B, Abdulnour RE, Rodman A, Manrai AK. Comparison of frontier open-source and proprietary large language models for complex diagnoses. JAMA Health Forum. 2025;6(3):e250040. doi:10.1001/jamahealthforum.2025.0040

Burke TA, Hamilton JL, Cohen JN, Stange JP, Alloy LB. Identifying a physical indicator of suicide risk: non-suicidal self-injury scars predict suicidal ideation and suicide attempts. Comprehensive Psychiatry. 2016;65:79–87. doi:10.1016/j.comppsych.2015.10.008

References (continued)

Care Quality Commission. Exploring evidence regarding vision-based monitoring in inpatient mental health units: rapid evidence review conducted by SQW. August 12, 2025.

Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. Science. 2024;383(6679):164–167. doi:10.1126/science.adg8538

Cheng HF, Stapleton L, Kawakami A, et al. How child welfare workers reduce racial disparities in algorithmic decisions. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. ACM; 2022:1–22. doi:10.1145/3491102.3501831

Chow DL, Miller SD, Seidel JA, Kane RT, Thornton JA, Andrews WP. The role of deliberate practice in the development of highly effective psychotherapists. Psychotherapy (Chic). 2015;52(3):337–345. doi:10.1037/pst0000015

Constantino MJ, Boswell JF, Coyne AE, Swales TP, Kraus DR. Effect of matching therapists to patients vs assignment as usual on adult psychotherapy outcomes: a randomized clinical trial. JAMA Psychiatry. 2021;78(9):960–969. doi:10.1001/jamapsychiatry.2021.1221

Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances. 2022;8(32):eabq6147. doi:10.1126/sciadv.abq6147

Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. Science. 1989;243(4899):1668–1674. doi:10.1126/science.2648573

Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine. 2021;27(10):1735–1743. doi:10.1038/s41591-021-01506-3

de Jong K, Conijn JM, Gallagher RAV, Reshetnikova AS, Heij M, Lutz MC. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. Clinical Psychology Review. 2021;85:102002. doi:10.1016/j.cpr.2021.102002

Delgadillo J, Ali S, Fleck K, et al. Stratified care vs stepped care for depression: a cluster randomized clinical trial. JAMA Psychiatry. 2022;79(2):101–108. doi:10.1001/jamapsychiatry.2021.3539

Delgadillo J, de Jong K, Lucock M, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. Lancet Psychiatry. 2018;5(7):564–572. doi:10.1016/S2215-0366(18)30162-7

References (continued)

DeRubeis RJ, Cohen ZD, Forand NR, Fournier JC, Gelfand LA, Lorenzo-Luaces L. The Personalized Advantage Index: translating research on prediction into individualized treatment recommendations. A demonstration. PLoS One. 2014;9(1):e83875. doi:10.1371/journal.pone.0083875

Dickerman BA, Hernán MA. Counterfactual prediction is not only for causal inference. European Journal of Epidemiology. 2020;35(7):615–617. doi:10.1007/s10654-020-00659-8

Freeman D, Lambe S, Kabir T, et al. Automated virtual reality therapy to treat agoraphobic avoidance and distress in patients with psychosis (gameChange): a multicentre, parallel-group, single-blind, randomised, controlled trial in England with mediation and moderation analyses. Lancet Psychiatry. 2022;9(5):375–388. doi:10.1016/S2215-0366(22)00060-8

Goldberg SB, Rousmaniere T, Miller SD, et al. Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. J Couns Psychol. 2016;63(1):1–11. doi:10.1037/cou0000131

Google. Google Workspace with Gemini FAQ [help center article]. Accessed November 9, 2026.

Greater Manchester Mental Health NHS Foundation Trust. VR therapy first piloted in Greater Manchester receives NICE early value assessment [news release]. December 19, 2023.

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: a meta-analysis. Psychological Assessment. 2000;12(1):19–30. doi:10.1037/1040-3590.12.1.19

Gustafson DH, McTavish FM, Chih MY, et al. A smartphone application to support recovery from alcoholism: a randomized clinical trial. JAMA Psychiatry. 2014;71(5):566–572. doi:10.1001/jamapsychiatry.2013.4642

Habicht J, Viswanathan S, Carrington B, Hauser TU, Harper R, Rollwage M. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. Nature Medicine. 2024;30(2):595–602. doi:10.1038/s41591-023-02766-x

He H, Li C, Ganglberger W, et al. What radio waves tell us about sleep! Sleep. 2025;48(1):zsae187. doi:10.1093/sleep/zsae187

Human Rights Data Analysis Group. How data analysis confirmed the bias in a family screening tool. June 22, 2023.

References (continued)

Jiang Y, Black KC, Geng G, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. NEJM AI. 2025;2(9). doi:10.1056/AIdbp2500144

Kasparov G. The chess master and the computer. The New York Review of Books. February 11, 2010.

Klasnja P, Smith S, Seewald NJ, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Annals of Behavioral Medicine. 2019;53(6):573–582. doi:10.1093/abm/kay067

Lemus HN, He H, Mirzazadeh A, et al. A wireless and contactless radio frequency sensor to detect bilateral tonic-clonic seizures and respiratory changes in the epilepsy monitoring unit: a phase 0/1 study. Epilepsy and Behavior. 2026;182:111108. doi:10.1016/j.yebeh.2026.111108

Marwaha JS, Choi J, Yuan W, Brat GA. The paradox of dissonant predictions: a central dilemma of physician-algorithm interaction. In: Proceedings of the Fifth Machine Learning for Health Symposium (ML4H 2025). PMLR. 2026;297:1632–1636.

Maybin S. Busting the attention span myth. BBC World Service. March 10, 2017.

McCarthy JF, Cooper SA, Dent KR, et al. Evaluation of the Recovery Engagement and Coordination for Health–Veterans Enhanced Treatment suicide risk modeling clinical program in the Veterans Health Administration. JAMA Network Open. 2021;4(10):e2129900. doi:10.1001/jamanetworkopen.2021.29900

Meehl PE. Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence. University of Minnesota Press; 1954. doi:10.1037/11281-000

Menger V, Spruit M, van Est R, Nap E, Scheepers F. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. JAMA Network Open. 2019;2(7):e196709. doi:10.1001/jamanetworkopen.2019.6709

Microsoft Canada. Attention spans [consumer insights report]. Spring 2015.

Microsoft. HIPAA and the HITECH Act [compliance offering documentation]. Accessed November 9, 2026.

References (continued)

Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Annals of Behavioral Medicine. 2018;52(6):446–462. doi:10.1007/s12160-016-9830-8

OpenAI. gpt-oss-120b model card. Hugging Face; 2025.

OpenAI. Introducing OpenAI for Healthcare. January 8, 2026.

Oxford Health NHS Foundation Trust. Oxevision camera-sensors and the care in your room [patient leaflet].

Richmond ME. Social Diagnosis. New York: Russell Sage Foundation; 1917.

Roske C, Nock MK. Clinical AND statistical prediction: integrating clinician judgment and statistical models to improve the prediction of suicidal behavior. Psychiatric Annals. 2025;55(12):e305–e309. doi:10.3928/00485713-20251024-02

Rueda A, Al-Shamali HF, Cote Z, et al. Enhancing psychiatry training using an agentic AI simulated consultation tool: prospective cohort study. JMIR Medical Education. 2026;12:e88580. doi:10.2196/88580

Russo J, Katon W, Zatzick D. The development of a population-based automated screening procedure for PTSD in acutely injured hospitalized trauma survivors. General Hospital Psychiatry. 2013;35(5):485–491. doi:10.1016/j.genhosppsych.2013.04.016

Serra-Garcia M, Gneezy U. Nonreplicable publications are cited more than replicable ones. Science Advances. 2021;7(21):eabd1705. doi:10.1126/sciadv.abd1705

Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence. 2023;5(1):46–57. doi:10.1038/s42256-022-00593-2

Shi YH, Tong CX, Xu W, et al. Predictive utility and associated challenges of adolescent arm scarring in suicide risk assessment. BMC Psychiatry. 2026;26(1). doi:10.1186/s12888-026-07968-4

References (continued)

Simon GE, Johnson E, Lawrence JM, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. American Journal of Psychiatry. 2018;175(10):951–960. doi:10.1176/appi.ajp.2018.17101167

Simon GE, Matarazzo BB, Walsh CG, et al. Reconciling statistical and clinicians' predictions of suicide risk. Psychiatric Services. 2021;72(5):555–562. doi:10.1176/appi.ps.202000214

Sperrin M, Jenkins D, Martin GP, Peek N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. Journal of the American Medical Informatics Association. 2019;26(12):1675–1676. doi:10.1093/jamia/ocz197

Stanley B, Brown GK, Brenner LA, et al. Comparison of the Safety Planning Intervention with follow-up vs usual care of suicidal patients treated in the emergency department. JAMA Psychiatry. 2018;75(9):894–900. doi:10.1001/jamapsychiatry.2018.1776

Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345–352. doi:10.1093/biostatistics/kxz041

U.S. Department of Health and Human Services. Guidance on HIPAA and cloud computing. Content last reviewed December 23, 2022.

University at Buffalo. UB study looks at the current state of ethically balancing AI and social work [news release]. August 14, 2026.

Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science. 2018;359(6380):1146–1151. doi:10.1126/science.aap9559

Wiest IC, Verhees FG, Ferber D, et al. Detection of suicidality from medical text using privacy-preserving large language models. British Journal of Psychiatry. 2024;225(6):532–537. doi:10.1192/bjp.2024.134

Yang Y, Yuan Y, Zhang G, et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. Nature Medicine. 2022;28(10):2207–2215. doi:10.1038/s41591-022-01932-x