Phoenix: International Multidisciplinary Research Journal
Vol 4, No. 2, April-June, 2026
ISSN: 2583-6897

ARTIFICIAL INTELLIGENCE AND COMPUTER SYSTEMS IN PSYCHOLOGICAL ASSESSMENT

Prof. (Dr.) Kranti Gawali
Professor and Head Dept.of Psychology Bhavan's College Autonomous Andheri west

Abstract
One big shift in how we study behavior comes from using smart machines along with computers when testing minds. Old ways of checking mental health used paper forms plus expert guesses, but now digital tools handle huge piles of actions and answers instead. What happens next often ties tech designs to trust in results, questions about steady outcomes, right-or-wrong choices, and where things might go later. Ideas mix together - from number-heavy test theory to pattern-spotting code - to weigh how well forecasts hold up under pressure. Hidden risks pop up too: hidden slants in software, private facts slipping out, decisions made behind closed doors. Even though speed, accuracy, and less hands-on work improve, success rests firmly on tight checks and fair rules guiding each step. Not humans, but helpers - that's how AI fits alongside psychologists, working together where needed. What comes next? Studies that dig into clear reasoning by machines, tracking behavior through data trails, testing algorithms across varied cultures - each step building on the last.

Artificial Intelligence Meets Psychological Assessment Through Machine Learning and Digital Psychometrics With Challenges in Algorithmic Bias and the Need for Explainable AI in Clinical Decision Support

1. Introduction

For years, psychological testing shaped how experts diagnose minds, judge learning needs, map workplace behaviors, and study thinking patterns. Back then, sessions leaned on fixed conversations, uniform surveys, plus watching actions - guided each time by tested measurement rules. Though such tools strengthened accuracy across studies, their reach stalls due to size hurdles, personal biases, slow results, along with poor live tracking options.

A shift in how computers work began changing things slowly. At first machines only handled grading tasks, making paperwork faster. Yet once smart software arrived - like programs that learn, think deeply, or understand speech - old ideas about measuring minds started crumbling. Instead of just scores, systems today study word choices, faces, voice shifts, body signals, even online actions to guess inner feelings more sharply.

Something shifted because many things pushed at once. One reason stands out - too few therapists face a growing number of people needing help, so new ways to measure mental states must stretch further. Another piece fits here - messages, posts, and online actions leave traces, patterns machines now learn to read. A third thread ties it together - spotting signs before crisis strikes becomes possible, opening paths where care acts sooner, not later.

Even with progress, some big issues still lack answers. Could algorithms measure mental traits as accurately as standard tests do? When predictions move beyond one group, might they fail elsewhere? Who takes responsibility if automated tools affect patient care choices? Exploring such matters means blending insights from tech, human behavior, moral rules, and policy design.

A look into how artificial intelligence blends with digital tools for mental health evaluation begins here. Machines that think like people meet tests meant to measure minds. Not just accuracy matters, but whether these methods stay fair across different lives. Questions pop up about privacy when data flows without clear rules. Scientists now chase new paths - some expected, others less obvious.

2. Foundations of Ideas and Thinking

2.1 Basic Ideas Behind Measuring Psychology

Figuring out how minds work starts with how we measure them. A common way splits a person's result into what they truly know and random mistakes. Trust in results comes from checking repeat answers, consistent scoring, and steady patterns inside tests. Even though many rely on this method, it treats every mistake the same - no matter who answers or which question shows up.

Instead of fixed formats, responses get shaped by hidden traits meeting question features. Because skills differ, tests adjust - accuracy grows when difficulty fits the person. Machines began tailoring questions once math models made real-time shifts possible.

Still, CTT and IRT came about before anyone dealt with massive behavioral datasets. Instead of relying on fixed hidden traits, modern AI tools find trends in complex data using probability methods. Because of this change, it is unclear whether old ideas about measurement accuracy still fit inside algorithm-based systems.

2.2 Artificial Intelligence Basics in Behavior Prediction

Machines that think? That's artificial intelligence - systems built to handle jobs needing human-style thinking. When it comes to testing minds, three kinds stand out

One way computers learn is by using examples that already have labels. These methods, like drawing lines between groups or building tree-like decisions, figure out emotional patterns from clear data. Another approach finds hidden groupings in behavior when there are no labels to start with. Patterns emerge simply by spotting similarities among actions taken over time.

Starting off with deep learning - it uses layered neural nets to spot patterns in voice, words, and images. Take CNNs - they handle reading emotions on faces quite well. When it comes to language or time-based data, RNNs come into play along with transformers that track sequence flow. Hidden layers stack up, making complex decisions possible without explicit programming.

2.3 Convergence: Digital Psychometrics

Out here, where mind meets machine, numbers start telling stories about people. Not just surveys anymore - what you do online spills secrets too: how fast fingers hit keys, delays between clicks, shifts in speech tones. Even posts tossed into social feeds carry clues. Behind each tap or pause, a hint hides - maybe stress, maybe openness, perhaps how hard the brain is working. Patterns form without anyone asking a single question. Data breathes where words fall short. Signals pile up quietly, building portraits no one sat down to draw. Built on shaky ground, turning actions into mental insights needs careful tuning. When predictions seem strong but ideas stay fuzzy, something's missing beneath the surface.

3. Computer-Based Psychological Testing Over Time

3.1 Computerized Adaptive Testing

Out here, computers first stepped into tests through adaptive systems. Depending on your last answer, the next question shifts - tightening time without losing accuracy. Schools began using these setups more and more. Not every method stuck like this one did.

3.2 Online and App Evaluations

Now available online, mental health tests show up on websites and phone apps. Though easier to reach, they still stick to old-style fixed question formats.

3.3 Ai Driven Systems Begin To Appear

This stage isn't just about going digital anymore - it's shifting into smart simulation. Machines now learn patterns, then predict outcomes using layered logic

- ·Analyze micro-expressions during video interviews
- ·Detect depressive speech markers through acoustic analysis
- ·Identify cognitive decline patterns via interaction metrics
- ·Predict suicide risk using linguistic and behavioral indicators

A change like this moves away from fixing problems after they happen, instead using data to guess what might come next.

4. AI Psychology Tools Using Computer Models

Fueled by artificial intelligence, mental health tools run on many kinds of digital frameworks.

4.1 Supervised Learning Models

One-way machines learn is through examples tagged with mental health labels. To guess diagnosis chances, some math formulas work step by step. What helps separate complex patterns? A method that stretches space between groups. When many small choices team up, results become steadier - trees do this together.

4.2 Deep Learning Architectures

Patterns hide in messy data, deep nets pull them out. Faces pass through CNNs, emotions show up. Time matters in how people speak - LSTMs track shifts that hint at mood troubles.

4.3 Multimodal Models

Something new is happening. Machines now pull together words, sounds, faces, maybe even heartbeats, just to guess what comes next a bit better. This blend isn't random. Feelings aren't flat - they twist and shift, shaped by many things at once. So the models grow wider, pulling in pieces that mirror how messy inner life really is.

4.4 Explainable AI (XAI)

It's hard to see inside deep learning models - they work like black boxes. To help, some approaches show how inputs shape outputs, pointing to key features along the way. When used in healthcare, being able to trace choices matters a lot - doctors must answer for decisions, and so must systems.

5. Measuring Psychological Accuracy in Artificial Intelligence

To work properly as tools for judging minds, artificial intelligence setups need to clear the bar set by older psychology tests. High prediction power on its own does not cut it. Even if a system sorts people correctly most times, it might miss what it claims to assess. So evaluation needs more than just number checks - grounding in how psychology measures traits matters too.

A single look at how AI behaves begins with whether it measures what it claims. Instead of just numbers, its design must reflect real concepts behind decisions made by machines. When results line up with outside evidence, trust grows without needing bold statements. Consistency appears not through repetition alone but across different settings and moments. What holds true here should hold elsewhere, even when conditions shift slightly.

5.1 Construct Validity

A test's value lies in how well it captures the mental idea it says it does. Instead of just assuming, old-school methods check this by rooting ideas in theory. Factor analysis comes into play, revealing patterns behind results. One measure links up with similar ones, showing alignment. At the same time, it stays distinct from unrelated concepts, proving separation.

When machines handle tasks involving human behavior, checking if they measure what they claim gets harder. Take a program that reads speech patterns showing sadness - it might spot more negative terms, frequent use of "I," or limited word choice. Yet spotting gloomy language does not mean it sees true mental health struggles. What looks like emotional weight could just reflect mood flavor, missing deeper issues like warped thinking, slowed movement, or trouble managing daily life.

5.2 Criterion Validity

When a model's results line up with trusted diagnostic tools, that’s what we mean by criterion validity. Matching today’s accepted methods shows concurrent validity, seen when new tests echo existing ones. How well it points to what happens later defines predictive validity, useful in guessing future behavior. The link between scores and real-world signs matters most here.

In AI-based systems, criterion validity typically involves comparing algorithmic predictions with:

- ·Structured clinical interviews
- ·DSM-5 or ICD-based diagnoses
- Common mental health assessment tools like PHQ-9 BDI and GAD-7
- ·Clinician-rated severity scores

A single example: when a system uses voice patterns to estimate chances of depression, what it returns should closely match scores from standard checklists or judgments made by doctors. Strong links shown through numbers like correlation values, along with good detection rates and low false alarms, help confirm it actually measures what it claims.

5.3 Reliability

One way to think about reliability? It's how steady a measure stays when repeated under similar circumstances. Sometimes it holds up well over days or settings - other times, less so. Picture giving the same test twice; if scores match closely, that’s a sign of dependability.

Researchers often check this by comparing results at different points in time. Another path involves looking inside the test itself - how its pieces fit together. A number called Cronbach's alpha sometimes shows whether items align consistently. When more than one person rates something, their level of agreement also matters. Matching judgments suggest the method works reliably.

Not every test fits how machines learn. Picking what matters means watching where mistakes happen. One wrong step changes everything. What works today might fail tomorrow. Watching patterns helps spot weak spots. Tools must shift as tasks grow harder. Results often surprise even experts

5.4 Generalizability

When a model works well only on certain data, its reach elsewhere becomes questionable. Picture city-dwellers speaking English shaping the system - then dropping it into villages where tongues mix and traditions differ; things might go sideways. Performance often stumbles outside familiar ground.

In AI-based psychological assessment, generalizability is particularly critical because:

- ·Cultural differences influence language use and emotional expression
- ·Socioeconomic context affects digital behavior patterns
- ·Gender and age variations alter communication styles

Might seem minor, yet skipping tests across different groups often leads to skewed results, wrong labels, unfair impacts. That's why checking performance outside the original data - using separate groups with varied backgrounds - matters most.

5.5 Evaluation Metrics and What They Miss

Most times, folks check how well AI works by looking at numbers that show its results, such as:

What counts as right guesses among all attempts shows how accurate the results are

- ·Precision: Proportion of predicted positives that are true positives
- What portion of true positive cases gets spotted accurately? That measure tells you recall, also called sensitivity
- ·F1-score: Harmonic mean of precision and recall
- ·ROC-AUC: Area under the receiver operating characteristic curve, reflecting discriminative capacity

Even though numbers can show how well predictions work, they miss whether the idea behind them makes sense. High scores on certain tests might come from patterns that aren't meaningful - just quirks in the data. Then again, when health records have very few examples of something serious, like suicide attempts, counting right answers could give a false impression.

So, numbers alone can't prove a test works well. Just because it looks good on paper doesn't mean it measures what it should. What matters comes down to how solid the method really is behind the scenes.

6.1 Algorithmic Bias and Fairness

Outcomes that favor certain demographics often stem from imbalanced learning inputs within artificial intelligence models. When it comes to mental health evaluations, distortions can surface due to limited sample diversity during development phases. Past disparities in clinical diagnosis sometimes carry forward into automated tools by design. Differences in language use or cultural context during information gathering further tilt results unintentionally.

Take, for example, models that detect depression through language. If these are mainly built using data from English speakers in Western countries, they might wrongly assess people from other cultures - whose ways of expressing emotional pain differ in speech patterns. In much the same way, software designed to interpret facial expressions can struggle with accuracy when applied to specific ethnic communities, especially if prior datasets underrepresent them.

6.2 Privacy and Data Protection

From voice samples to online posts, digital footprints feed machine-based mental health evaluations. Though collected passively, these traces may expose intimate patterns unnoticed even in therapy rooms. While typing rhythms or sleep cycles seem ordinary alone, combined they sketch private emotional landscapes. Because smartphones track movement and messaging alike, insight emerges where users might expect none. Even heartbeat variations recorded by watches carry clues once only verbal reports could provide. When algorithms piece together such fragments, self-revelation occurs without intent. What people share freely online gains meaning when viewed alongside body signals captured daily. Such methods draw conclusions not from interviews - but from accumulated moments lived publicly yet quietly. Hidden tendencies surface through repetition, timing, context - elements rarely discussed aloud. Without asking, systems infer mood shifts using data considered mundane elsewhere. Quiet behaviors, stitched across weeks, become evidence of inner states once known solely through confession. The depth arises less from single points than their steady accumulation over time.

Privacy risks include:

- ·Unauthorized data access or breaches
- ·Secondary data usage without consent
- ·Re-identification of anonymized datasets
- ·Surveillance concerns in organizational contexts

Because mental health information demands careful handling, strong oversight of data practices becomes necessary. For responsible implementation, clear rules must guide how systems are used

- ·Strong encryption and cybersecurity protocols
- ·De-identification and anonymization procedures
- ·Strict data minimization policies
- ·Clear limitations on data retention

Meeting rules like those guarding health information and general data laws isn’t optional. Still, just following the law doesn’t mean actions are right - firms need forward-thinking approaches that build privacy into how systems work from the start.

6.3 Informed Consent and Autonomy

Besides outlining goals, methods, dangers, and privacy rules, standard psychological evaluation requires clear agreement from participants. Since artificial intelligence often works invisibly, changes without notice, and runs within apps where people might not realize testing is happening, understanding what one agrees to becomes harder.

True informed consent in AI-driven assessment must address:

What kind of information is gathered - both given directly and recorded without input - affects how it's used

- ·The purpose of algorithmic analysis
- ·Potential risks and limitations
- ·Data sharing policies
- ·The extent of automation in decision-making

Power imbalances often shape decisions around agreement, especially where authority figures are involved. When people feel pressured, true willingness fades. Protection of personal choice depends on clear safeguards within ethical systems. Opting out should carry no negative consequences. Understanding how algorithms affect outcomes is essential for meaningful participation. Clarity about automated processes supports informed judgment.

6.4 Clinical Responsibility and Legal Duty

Should an AI miss signs of emotional distress, blame might stretch beyond the software creator. A mistake in forecasting danger could involve clinicians who used the tool, developers who designed it, even institutions that approved its use. Where fault lies grows unclear when technology shapes mental health judgments. Errors spark debate about oversight, especially if someone is wrongly labeled at risk. Responsibility shifts depending on who trusted the system, how they applied it, later

- ·Software developers
- ·Data scientists
- ·Clinical institutions
- ·Individual practitioners

Ambiguity around who is responsible creates challenges in both law and ethics. When automation is used too heavily, it can weaken a practitioner's sense of duty; when used too little, valuable warnings might be missed.

6.5 Transparency in the Face of Opaque Systems

Although powerful, numerous artificial intelligence setups like deep learning models often work behind a veil of secrecy, delivering results with little clarity about how decisions emerge. Because mental health evaluations demand transparency, such hidden processes weaken confidence, challenge expert understanding, sometimes even erode fairness in practice.

Understanding how risk scores emerge matters for clinicians using them in diagnosis. Since judgments about a person's mind can shape care, patients often seek clarity on how such decisions form. Not knowing the process behind results can weaken trust. When predictions influence treatment paths, seeing the logic helps both doctor and patient move forward.

7. Applications Across Psychological Domains

Far beyond just one area, artificial intelligence shapes how mental health evaluations unfold across varied fields. From schools to hospitals, workplaces to brain-function studies, digital tools adapt differently each time. Where methods must shift, so do rules around fairness, accuracy, trust. Challenges pop up - not only in design but also in proving these systems actually work as intended.

7.1 Clinical Psychology

Clinicians now rely more on artificial intelligence to spot early signs of mental health issues, track symptoms over time, one benefit being improved forecasts of patient risks. Instead of traditional methods, machine learning digs into how patients describe their experiences, searching for subtle clues tied to conditions like depression, anxiety, or PTSD. Take voice analysis: certain shifts in tone, pauses between words, even word choice may signal emotional distress. Lower variation in vocal pitch, longer gaps when speaking, negative themes in language - these often surface in people dealing with depression. What stands out is how consistently these patterns appear across different cases, supporting the role of such tools in everyday practice.

7.2 Educational Psychology

Starting mid-stream, computer-driven quizzes track how students think, pulling data on understanding as they go. Instead of fixed questions, these tools shift challenge levels during tests, matching each learner's pace. Not limited to scores alone, artificial intelligence examines attention clues - like hesitation times or repeated attempts - to gauge effort and mental strain. From behind the scenes, subtle behaviors shape insights into both motivation and thinking demands.

7.3 Workplace Behavior and Job Performance

When companies hire new staff, they often turn to artificial intelligence to judge personality traits. Instead of relying only on resumes, some systems study how candidates speak during video interviews. These programs detect subtle facial movements that might signal confidence or hesitation. Rather than asking direct questions, software examines word choice across answers. Traits like reliability or sociability get estimated through repeated behavioral signals. Even small shifts in voice tone may influence results. Leadership potential is sometimes judged by matching responses to past successful managers. Engagement levels come into view when employees interact with digital platforms over time.

7.4 Brain Mind Thinking Learning

Facing complex cases, neuropsychologists once depended solely on structured test sets given face-to-face. Today, subtle cues - like how quickly someone responds or where mistakes cluster - are pulled from digital behavior and interpreted using artificial intelligence.

Spotting signs of brain-related decline - like mild memory issues or dementia - is advancing quickly. Though small, shifts in how someone types, pauses while speaking, or moves through online spaces might signal early risk. Because these behaviors are tracked over time, computers learn what normal looks like for each person. When actions drift too far from that pattern, it could point to mental changes beginning beneath the surface. Models built using deep learning detect those differences before obvious symptoms appear.

8. Traditional and AI Psychological Assessments Compared

A comparative examination of traditional psychological assessments and AI-driven systems reveals both continuity and divergence in methodological philosophy. Conventional psychometric instruments are grounded in decades of empirical validation, theoretical clarity, and standardized administration procedures. Their strength lies in conceptual transparency and well-established norms, allowing clinicians to interpret scores within clearly defined frameworks of reliability and validity. In contrast, AI-based systems excel in managing high-dimensional data and identifying nonlinear patterns that may be imperceptible through classical statistical approaches. Machine learning algorithms can process linguistic nuances, behavioral micro-signals, and large-scale longitudinal datasets with remarkable precision. However, predictive strength does not necessarily equate to explanatory depth. A model may achieve high classification accuracy while offering limited theoretical insight into the psychological mechanisms underlying its predictions. Thus, the contrast is not merely technical but epistemological between measurement anchored in theory and prediction optimized through data-driven computation.

9. Technical Challenges and Constraints

Few tools work well right now because built-in flaws limit how AI measures mental health.

9.1 Data Quality and Hidden Biases in Labels

Reliability of machine learning depends heavily on the quality of its training data. When labels are missing, populations lack diversity, or samples misrepresent reality, predictions tend to drift from accuracy. Subjective decisions by clinicians during annotation may further complicate consistency in mental health datasets.

9.2 Overfitting and Changes in Data

When models learn from limited examples, their predictions can falter elsewhere. Real-world conditions often drift away from original assumptions - this is dataset shift. Because of such changes, ongoing testing becomes necessary. Retraining helps maintain accuracy over time.

9.3 Infrastructure and Accessibility

Where resources are limited, setting up artificial intelligence often fails due to missing tech foundations. Without steady internet, cloud platforms cannot function properly. Heavy computing needs go unmet when processing power is too weak.

10. Human–AI Collaboration Framework

One path ahead for psychology's evaluation methods blends algorithms alongside trained professionals. Machine learning might support clinicians instead of replacing them. Thinking about tools that assist rather than dominate could shift how assessments evolve. Experts may guide systems while data sharpens decisions. This mix balances precision with personal insight. Not either machines or people - but both shaping outcomes differently. How these pieces interact matters more than which one leads.

10.1 Decision-Support Systems

Starting with a probability score, artificial intelligence offers preliminary assessment cues for closer review. Rather than replacing judgment, visual summaries in support systems break down influencing factors clearly. A dashboard might show which elements weigh more heavily, helping medical staff grasp reasoning behind alerts. Sometimes uncertainty remains, yet transparency builds trust gradually. Early signals do not confirm conditions - instead they suggest where attention could go next.

10.2 Augmentation Model

Still, within augmentation frameworks, artificial intelligence sharpens speed and spotting of patterns - yet leaves judgment to people. Responsibility for weaving insights into context, building trust in therapy, and upholding ethics stays firmly with the clinician.

10.3 Professional Training Implications

When artificial intelligence enters evaluation methods, clinicians must learn how to work with digital tools. Grasping how algorithms operate comes before spotting biases in results. Handling patient information responsibly matters just as much as interpreting outcomes. Skills in these areas will shape future readiness for practitioners.

11. Future Directions

The trajectory of AI in psychological assessment suggests several emerging pathways.

11.1 Digital Phenotyping

From moment to moment, behavior tracked through phones or wearables supports long-term observation of psychological states. Without active input, these automatic recordings can detect subtle shifts away from usual patterns.

11.2 Clear and Understandable Artificial Intelligence

What drives progress in explainable AI is a focus on clarity, so clinicians and patients feel more confident in its use. While transparency becomes clearer, reliance grows - not just by experts but also those receiving care.

11.3 Cross-Cultural Validation

Working across regions means training data needs awareness of cultural context, while checks during development should reflect that too. Where languages differ, machine understanding must adapt to subtle differences in expression.

11.4 Brain–Computer Interfaces (BCI)

Few studies now examine how brain signals might link straight into artificial intelligence systems - opening paths toward accurate detection of mental states. Though still early, such work suggests machines could one day interpret thought patterns with close precision.

11.5 Custom Models for Mental Health

Future systems may generate individualized psychological profiles using adaptive algorithms that evolve with user behavior.

12. Conclusion

Fueled by algorithms, artificial intelligence is shifting how psychological evaluations are designed and applied. Moving past conventional methods, computerized systems allow broader reach through automated analysis of behavioral patterns. While these tools offer forecasting abilities, their deployment demands careful scrutiny alongside moral considerations. Human judgment remains essential even when machines assist decision-making.

What keeps AI useful in psychology isn't replacement, but its role alongside human insight. Lasting integration hinges not only on technology, yet equally on ongoing conversation across fields. Clarity in how models work matters as much as testing them beyond Western contexts. Ethical practice forms the base - without it, even advanced tools lose grounding.

Beyond just tech progress, using computational intelligence in psychology changes entirely how we gauge and make sense of thoughts and actions. Instead, measurement now leans on data-driven insight rather than traditional observation alone. With machines involved, interpretation shifts subtly - less intuition, more pattern recognition. Understanding mental states becomes less about subjective judgment, more about modeled outcomes. This turn alters foundational methods once taken for granted in clinical practice.

References

1. Abd-Alrazaq, A. A., Alhuwail, D., Househ, M., Hamdi, M., Alhajj, M., & Shah, Z. (2022). The performance of artificial intelligence-driven technologies in diagnosing mental disorders: An umbrella review. *NPJ Digital Medicine, 5*, Article 87. https://doi.org/10.1038/s41746-022-00631-8.
2. Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: A narrative review. *NPJ Digital Medicine, 5*(1), 1–13. https://doi.org/10.1038/s41746-022-00589-7.
3. Kerz, E., Zanwar, S. Z., Qiao, Y., Wiechmann, D. W., & others (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in Psychiatry, 14*, Article 1219479. https://doi.org/10.3389/fpsyt.2023.1219479.
4. González, O. (2020). Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behavioral Research, 56*(6), 903–919. https://doi.org/10.1080/00273171.2020.1781585.
5. Kline, A., Kline, T., Hossein Abad, Z. S., & Lee, J. (2020). Using item response theory for explainable machine learning in predicting mortality in the intensive care unit: A case-based approach. *Journal of Medical Internet Research, 22*(9), e20268. https://doi.org/10.2196/20268.
6. Ehtemam, H., Sadeghi Esfahlani, S., Sanaei, A., Ghaemi, M. M., Hajesmaeel-Gohari, S., Rahimisadegh, R., Bahaadinbeigy, K., Ghasemian, F., & Shirvani, H. (2024). Role of machine learning algorithms in suicide risk prediction: A systematic review-meta-analysis of clinical studies. *BMC Medical Informatics and Decision Making, 24*(1), 138. https://doi.org/10.1186/s12911-024-02524-0.
7. Su, C., et al. (2020). Machine learning for suicide risk prediction in children and adolescents with electronic health records. *Translational Psychiatry, 10*(1), 413. https://doi.org/10.1038/s41398-020-01100-0.
8. Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffman, B., & others (2020). Validation of an electronic health record–based suicide risk prediction modeling approach across

multiple health care systems. *JAMA Network Open, 3*(3), e201262. https://doi.org/10.1001/jamanetworkopen.2020.1262.

9. Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.
10. Greco, A., Gagliardi, C., & others (2021). (Representative empirical articles on transformer/NLP performance for depression detection — see Zhang, 2022 for synthesis and dataset references.) *NPJ Digital Medicine* and other journals (see Zhang, 2022). https://doi.org/10.1038/s41746-022-00589-7.
11. Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2018). New tools for new research in psychiatry: A scalable and privacy-preserving approach to digital phenotyping. *npj Digital Medicine* (see related digital-phenotyping literature; Insel 2018 commentary). https://doi.org/10.1038/s41746-018-0029-3.
12. Insel, T. R. (2018). Digital phenotyping: A global tool for psychiatry. *World Psychiatry, 17*(3), 276–277. https://doi.org/10.1002/wps.20550.
13. Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnousi, F. (2020). Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. *International Journal of Environmental Research and Public Health, 17*(16), 5929. https://doi.org/10.3390/ijerph17165929.
14. (Psychometrics + ML methodological bridge) Martínez-Plumed, F., Prudêncio, R., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in AI: analysing machine learning classifiers at the instance level. *Artificial Intelligence*, *271*, 18–42. https://doi.org/10.1016/j.artint.2018.12.005.