

The Psychiatric Examination

From Practice to Evidence

Husain Alkhalidy · Department of Internal Medicine, King Khalid University, Abha · September 2026

Part 1 — The examination

Setting up. In psychiatry the interview is the examination. There is no organ to expose and no stethoscope to place; what is examined is the patient's appearance, speech, mood, thought, perception and cognition as they are shown during a conversation that has been arranged to show them. The examination therefore begins in the waiting room and in the corridor, continues through the history, and is written up afterwards under fixed headings — the **mental state examination** — so that a second clinician can reconstruct what was seen and heard rather than what was concluded. Three things are settled before the first question: safety, for the patient and the examiner, with the door, the seating and a colleague considered; privacy, with the family asked to wait and then asked to speak; and the collateral history, which in psychiatry is not an optional extra but half the evidence. And the physical examination is not skipped because the complaint is psychological — the thyroid, the pupils, the tremor, the gait, the scars on the forearms, the waist and the blood pressure, and the neurological screen that finds the organic cause are all part of it.

Appearance and behaviour

- **Appearance.** Age against stated age; dress, grooming and hygiene; weight and evidence of its change; the marks on the body — self-harm scars on the forearms and thighs, injection sites, tattoos, bruising; the smell of alcohol; the pupils.
- **Behaviour and rapport.** Eye contact, facial expression, posture; whether rapport is established, guarded, hostile or over-familiar; distractibility and responding to unseen stimuli; the level of arousal.
- **Psychomotor activity.** Retardation — the slowed movement, long latencies and reduced gesture of severe depression — or agitation, restlessness and pacing; the akathisia of antipsychotics, distinguished by the patient's subjective restlessness; tremor; the tics, mannerisms and stereotypies.
- **Abnormal movements.** Tardive dyskinesia of the mouth, tongue and fingers, scored on the Abnormal Involuntary Movement Scale; parkinsonism; dystonia; and the signs of catatonia — immobility, mutism, staring, posturing, waxy flexibility, negativism, echolalia and echopraxia, stereotypy, verbigeration, withdrawal — elicited with the standardised Bush–Francis examination rather than noticed in passing.
- **Intoxication and withdrawal.** The smell, the slurring, the nystagmus and ataxia of alcohol; the pinpoint pupils of opioids; the dilated pupils, sweating and tachycardia of stimulants; the tremor, sweating, tachycardia, anxiety and hallucinosis of alcohol withdrawal, scored on the Clinical Institute Withdrawal Assessment.

Speech

Rate, volume, quantity, latency and prosody: the pressured, loud, unstoppable speech of mania; the slow, quiet, monosyllabic speech with long pauses of depression; the poverty of speech of negative-symptom schizophrenia; the flat prosody; the dysarthria that is a neurological finding and the dysphasia the neurological chapter separated from it. A sample is recorded verbatim, because the form of thought is inferred from it and a paraphrase destroys the evidence.

Mood and affect

Mood is what the patient reports — asked in their own words and rated by them on a scale of ten — with its diurnal variation, its reactivity to events, and the biological accompaniments asked for directly: sleep and its pattern of waking, appetite and weight, energy, libido, concentration. **Affect** is what the examiner observes: its range (full, restricted, blunted, flat), its reactivity, its lability, and its congruence with the mood described and the content discussed. An incongruous affect — laughter while describing a loss — and a discrepancy between reported mood and observed affect are findings in their own right.

Thought: form and content

- **Form.** Whether the thought is coherent and goal-directed, and if not, how it fails: circumstantiality that arrives eventually, tangentiality that never does, flight of ideas with its clang associations and pressure, loosening of associations and derailment, thought block, neologisms, perseveration, word salad. The verbatim sample is the evidence.
- **Content.** Preoccupations; obsessions, recognised by the patient as their own and resisted; overvalued ideas; and delusions, for each of which the type (persecutory, grandiose, referential, nihilistic, of guilt, jealousy, love, infestation, control), the degree of conviction, the fixity against evidence, the systematisation, and the congruence with mood are recorded. The passivity phenomena — thought insertion, withdrawal and broadcast, made feelings, impulses and actions — are asked for by name.
- **Risk to self.** Suicidal ideation is asked about directly, in plain words, in every examination; asking does not plant the idea. It is then graded: passive wish to be dead, active ideation, plan, intent, means and their availability, preparation and rehearsal, previous attempts and their lethality, and the protective factors — dependants, faith, plans for the future. Self-neglect and self-harm without suicidal intent are recorded separately.
- **Risk to others.** Thoughts of harming others, their targets, plans and means; the command hallucinations that direct them; the past history of violence, which is the strongest single predictor.

Perception

Hallucinations by modality, asked for directly and observed indirectly in the patient who pauses, turns or answers an unheard voice: auditory — second-person, or the third-person voices, running commentary and thought echo that carry particular weight; visual, which point first to delirium,

drugs and neurological disease; olfactory, gustatory and tactile. Illusions; depersonalisation and derealisation; the pseudo-hallucinations experienced within the mind and recognised as such.

Cognition

Orientation in time, place and person; attention, by the months of the year backwards or serial sevens; registration and recall; and then the instruments the neurological chapter set out — the 4AT or the Confusion Assessment Method when delirium is possible, and the Montreal Cognitive Assessment or the Mini-Mental State Examination for cognition, read against the patient's education and language. In the psychiatric examination the point of the cognitive screen is the one diagnosis that changes everything: the delirium presenting as psychosis, the dementia presenting as depression, and the depression presenting as dementia.

Insight, judgement and capacity

- **Insight**, graded rather than recorded as present or absent: awareness that something is wrong; attribution of it to illness; acceptance that treatment is needed; willingness to take it.
- **Judgement**, as the decisions the patient has made and would make — the money spent, the risks taken, the response to a hypothetical.
- **Capacity**, which is decision-specific and time-specific and is not the same as any of the above: the ability to understand the information relevant to the decision, to retain it, to weigh it, and to communicate a choice, assessed formally — with the Aid to Capacity Evaluation where the answer will be acted on — and never inferred from a diagnosis, a cognitive score or a refusal.

The instruments

The mental state examination is a description; the instruments are the measurements laid over it, and each belongs at a named point. The **PHQ-9** for depression and the **GAD-7** for anxiety at the end of the history of mood; the **CAGE** questions or the **AUDIT-C** in the substance history; the **4AT** in cognition; the **Columbia** scale or an equivalent structured inquiry for suicidal ideation, used to ensure the questions are asked and recorded, not to predict; the **Bush–Francis** examination for catatonia, the **Abnormal Involuntary Movement Scale** for dyskinesia and the **CIWA-Ar** for withdrawal in behaviour; and the **Aid to Capacity Evaluation** when capacity is in question.

The physical examination that belongs to psychiatry

The thyroid and the signs of its status; the pupils, the tremor, the gait and the reflexes; the waist circumference, blood pressure, and the metabolic syndrome that antipsychotics produce; the extrapyramidal examination; the forearms and thighs; the signs of intoxication and withdrawal; the fundi and the neurological screen in any first psychosis, any psychosis after forty, and any presentation with visual hallucinations, fluctuating consciousness or focal signs. The organic cause is excluded by examination, not by the absence of complaint.

Putting the signs together

The mental state examination sorts presentations by their pattern across the headings, and the pattern — not the single symptom — is the syndrome.

Pattern	The findings that make it
Depressive episode	Psychomotor retardation or agitation; slow, quiet speech with long latencies; low mood with reduced reactivity and restricted affect; diurnal variation with early waking; guilt, worthlessness and hopelessness; suicidal ideation; impaired concentration; mood-congruent delusions and hallucinations when psychotic.
Mania	Overactivity, disinhibition and reduced sleep; pressured speech; elevated or irritable mood with labile affect; flight of ideas; grandiose delusions; distractibility; poor insight and judgement.
Schizophrenia and acute psychosis	Self-neglect, guardedness, responding to voices; poverty or disorder of speech; blunted or incongruous affect; loosening of associations; persecutory and referential delusions and passivity phenomena; third-person auditory hallucinations; intact orientation; impaired insight.
Delirium mistaken for psychosis	Fluctuating level of consciousness; inattention on the months backwards; disorientation; visual hallucinations; a physical cause on examination — the pattern that separates it from every functional psychosis.
Catatonia	Immobility or excitement, mutism, staring, posturing, waxy flexibility, negativism, echophenomena and stereotypy on the Bush–Francis examination; autonomic instability when malignant.
Anxiety and panic	Restlessness, tremor and sweating; rapid speech; anxious affect with preserved reactivity; worry as the content, with panic attacks, avoidance and hypervigilance; normal cognition; a thyroid, a pulse and a substance history examined.
Intoxication and withdrawal	The pupils, the smell, the tremor and the vital signs; visual and tactile hallucinations; a fluctuating course; the CIWA-Ar score in alcohol withdrawal.
Depression presenting as dementia	'I don't know' answers rather than wrong ones; a depressed affect that precedes the cognitive complaint; variable performance; improvement with treatment.
Eating disorder	Low weight and the signs of starvation — bradycardia, hypotension, lanugo, cold peripheries — or a normal weight with parotid swelling, dental erosion and Russell's sign; a distorted body image as overvalued idea; the electrolytes examined.
Functional neurological disorder	The neurological chapter's positive signs — Hoover's sign, drift without pronation — in a patient whose mental state examination carries the stressor and the affect.

Part 2 — What the literature says

How well do examiners agree?

Psychiatry measured the reliability of its own examination earlier and more thoroughly than any other specialty, because its diagnoses had no reference standard beyond the examiner, and the field

trials for the fifth Diagnostic and Statistical Manual are the clearest statement of what that measurement found. Two clinicians interviewing the same patient on separate occasions, with usual clinical methods, agreed on major depressive disorder with a kappa of about 0.3 and on generalised anxiety disorder at 0.2 — both in the range the trials called questionable — while post-traumatic stress disorder reached 0.67, schizophrenia 0.46, and bipolar I disorder and alcohol use disorder sat in the good range of 0.40 to 0.59; mixed anxiety-depressive disorder fell below 0.20 and was dropped. Thirty-nine per cent of the diagnoses tested were questionable or unacceptable [1, 2]. The signs that are elicited by a standardised examination do far better: the Bush–Francis catatonia scale reproduces between raters at 0.93 to 0.96 and its screening instrument at 0.95 to 0.97 [3, 4]; the withdrawal scale's inter-rater reliability is robust [5]; and a video link changes the diagnosis less than the diagnosis itself varies, with concordance between telepsychiatry and face-to-face assessment at kappa 0.82 across sixteen disorders in one analysis and 0.46 with heavy heterogeneity in another [6]. The examiner's most consequential single judgement — whether the patient has capacity — is the one most often missed: capacity is lacking in 26 per cent of medical inpatients, and clinicians recognise it in 42 per cent of them [7].

Judgement or instrument	Agreement	Comment
Major depressive disorder (DSM-5 field trials)	$\kappa \approx 0.3$	Questionable; usual clinical interview, test–retest [1]
Generalised anxiety disorder	$\kappa \approx 0.2$	Questionable [1]
Post-traumatic stress disorder · schizophrenia	$\kappa 0.67 \cdot 0.46$	Very good · good [1]
Bipolar I · alcohol use disorder	$\kappa 0.40 - 0.59$	Good [1]
Mixed anxiety-depressive disorder	$\kappa < 0.20$	Unacceptable; not adopted [1]
Bush–Francis catatonia rating scale · screening instrument	ICC 0.93 – 0.96 · 0.95 – 0.97	A standardised elicited examination [3, 4]
CIWA-Ar	robust; excellent ICC	Scored, not impressionistic [5]
Telepsychiatry against face-to-face diagnosis	$\kappa 0.82$ (16 disorders) · 0.46 (pooled, I^2 92 per cent)	The medium changes less than the examiners do [6]
Recognition of incapacity by treating clinicians	42 per cent of incapacitated patients	Accurate when they do recognise it; prevalence 26 per cent [7]
Large language models against clinicians on psychopathology	0.81 · 0.76 · 0.60 vs 0.79 · 0.68 · 0.58	Depression · mania · schizophrenia; the model and the clinicians fail on the same cases [8]

How accurate are the instruments? The Rational Clinical Examination series and the meta-analyses

Question	Instrument	Likelihood ratio or accuracy	Source
Is this patient clinically depressed?	Case-finding questionnaires, pooled	Median LR+ 3.3 (range 2.3 – 12.2), median LR– 0.19 (0.14 – 0.35)	Williams 2002 [9]
	PHQ-9 at 10 or above	Sensitivity 88 per cent, specificity 85 per cent against a semi-structured interview; 58 studies, 17,357 participants	Levis 2019 [10]
Is this patient anxious?	GAD-7 at 8 or above	Sensitivity 92 per cent, specificity 76 per cent	Plummer 2016 [11]
Does this patient have an alcohol problem?	CAGE questions	The likelihood ratio rises steeply with each positive answer, from well below 1 at a score of 0 to double figures at 3 and 4; the AUDIT-C is more sensitive for hazardous drinking short of dependence	Kitchens 1994 [12]; Buchsbaum 1991 [13]; Bush 1998 [14]
Does this patient have decision-making capacity?	Aid to Capacity Evaluation	LR+ 8.5, LR– 0.21	Sessums 2011 [7]
	MMSE below 20	LR+ 6.3 for incapacity; useful only at the extremes	
Will this patient die by suicide?	Any risk scale or stratification, pooled	Positive predictive value 5.5 per cent for suicide, 26 per cent for self-harm; no instrument accurate enough to allocate treatment	Carter 2017 [15]
	Fifty years of risk-factor research	Slight predictive power for ideation, attempts and deaths, not improving over time	Franklin 2017 [16]
	Machine-learning prediction models	Positive predictive values 6 – 17 per cent in-sample and as low as 0.1 per cent in low-prevalence populations; the base rate defeats the model	Belsher 2019 [17]; meta-analyses 2022 – 2025 [18, 19]
Does asking about suicide cause harm?	Direct inquiry	No increase in ideation in 13 studies; possible reduction in treatment-seeking populations	Dazzi 2014 [20]
Will this patient be violent?	Structured risk instruments	Better at ruling out than ruling in; a high-risk classification is wrong in most patients	Fazel 2012 [21]

Question	Instrument	Likelihood ratio or accuracy	Source
Is this patient delirious?	4AT · Confusion Assessment Method	See the neurological chapter: CAM LR+ 9.6, LR– 0.16; 4AT pooled sensitivity and specificity 88 per cent	Wong 2010; Tieges 2021
Can a machine hear depression?	Automatic speech analysis, 105 studies	Best reported: sensitivity 0.84, specificity 0.83; worst reported: 0.63 and 0.60 — a complementary method, not a standalone test	Meta-analysis 2025 [22]
	Language-based machine learning, 28 studies	Precision 0.78, recall 0.76, area under the curve 0.79	Meta-analysis 2026 [23]
	Wearable-sensor AI	Sensitivity 0.87 – 0.89, specificity 0.91 – 0.93 in the studies pooled	Meta-analyses 2023 – 2026 [24]
	Smartphone digital phenotyping, 24 studies	Moderate performance; missing data and no external validation	Systematic review 2024 [25]

Three things stand out. **The questionnaires are better than the diagnosis they screen for:** a nine-item form agrees with a structured interview more closely than two psychiatrists agree with each other about the same illness, which is why the form has become the examination's spine rather than its supplement. **Risk cannot be predicted at the bedside and the literature has stopped pretending otherwise:** a scale that calls a patient high risk is wrong nineteen times in twenty for suicide, and fifty years of risk factors and a decade of machine learning have not moved that number, because the base rate will not let them; the honest use of a suicide inquiry is to make sure the questions are asked and the answers acted on, not to sort patients. And **the single judgement that clinicians make wrongly most often is capacity**, which they miss in six of ten patients who lack it while an instrument that takes ten minutes has a positive likelihood ratio of eight.

The examiner is the limiting reagent

The field trials are a study of examiners. The same manual, the same patients and the same training produced a kappa of 0.3 for the commonest diagnosis in medicine, and the trials' own authors located the problem in the unstructured clinical interview rather than the criteria [1, 2]. The capacity data say the same thing from the other side: the physicians who assessed their patients informally missed most of the incapacity present, and when they used an instrument they did not [7]. And the comparison that arrived in 2026 — a large language model matched against practising clinicians on standardised psychopathology, equal on depression, ahead on mania and level on schizophrenia — is less a claim about machines than a measurement of how much of the examination is pattern recognition that a structured process can reproduce [8]. The catatonia scale, the withdrawal scale and the screening questionnaires are that structured process applied to the bedside.

Technique changes the answer

- **Ask about suicide directly, every time, in plain words**, and grade the answer; the evidence that asking harms does not exist, and the evidence that not asking misses is the whole literature [20].
- **Record speech verbatim** and thought form from the sample; a paraphrase is the examiner's diagnosis, not the patient's examination.
- **Use the questionnaire at the end of the history**, not instead of it: the PHQ-9 and GAD-7 have known thresholds and likelihood ratios; the impression does not [9, 10, 11].
- **Elicit catatonia and dyskinesia with the standardised examinations**, which reproduce at 0.93 and above; noticed in passing, they are missed [3, 4].
- **Assess capacity as a decision, not a diagnosis**: understand, retain, weigh, communicate, with the Aid to Capacity Evaluation when it will be acted on, and never from the cognitive score alone except at its extremes [7].
- **Treat the risk assessment as an inquiry and a plan, not a prediction**; document what was asked, what was answered and what was done [15, 16].
- **Examine the body**: the thyroid, the pupils, the tremor, the gait, the waist and the fundi, in every first presentation; the organic psychosis is found by the examination the psychiatric history tempts the examiner to omit.
- **Take the collateral history**, and count it as examination.

What has changed, 2020 – 2026

The examination went through a camera and survived. The pandemic made telepsychiatry the default, and the 2025 meta-analysis of diagnosis and symptom assessment across settings found concordance from moderate to almost perfect depending on the disorder and the scale, with the caveat that the pooled estimate hides wide variation and that the elicited signs — the tremor, the gait, the subtle affect — travel worst [6]. The mental state examination, being mostly observation and conversation, moved online more easily than any examination in this series.

The machine listens, reads and wears. A 2025 meta-analysis of 105 studies of automatic speech analysis for depression found the best-reported models at sensitivity 0.84 and specificity 0.83 and the worst at 0.63 and 0.60, and concluded that the method is complementary rather than standalone [22]; language-based classifiers pool to an area under the curve of 0.79 [23]; wearable-sensor models report sensitivities near 0.9 in the studies that exist [24]; and digital phenotyping from the smartphone, reviewed across 24 studies, reaches moderate performance with the recurring problems of missing data and no external validation [25]. Natural-language markers of psychosis risk are being pursued with the warning that language is a social marker before it is a biological one [26]. None of these has entered routine examination; all of them are measuring what the mental state examination describes in words — rate, prosody, coherence, movement, sleep.

The large language model took the examination. Tested on standardised psychiatric knowledge, the general models encode it accurately [27]; tested on real multicentre clinical records, they diagnose

with accuracy that varies by condition and falls on early schizophrenia [28, 29]; and benchmarked against practising clinicians on psychopathological assessment in 2026, the current model matched or exceeded them on depression and mania and matched them on schizophrenia — which was also where both did worst [8]. The question the field is now asking is not whether the model can do the examination but what the examination is for when it can.

Risk prediction was declared a base-rate problem. The machine-learning literature that promised to succeed where the scales failed was pooled repeatedly between 2019 and 2025 and found the same positive predictive values the scales had — single figures in low-prevalence settings — because no classifier escapes the arithmetic of a rare outcome [17, 18, 19]. The 2025 adolescent meta-analysis and the 2024 review of models in psychiatric populations both end where Carter and Franklin ended: assess to act, not to predict [15, 16, 19].

Part 3 — Practical synthesis for teaching

- Teach the mental state examination as a description under fixed headings, written so that a colleague can see what was seen; teach students to quote the patient and to record the affect they observed beside the mood they were told.
- Teach the instruments as the measurement layer of the examination, each at its named point, and teach their numbers: a PHQ-9 of 10 carries a sensitivity of 88 per cent and a specificity of 85; a positive CAGE answer multiplies the odds; a 4AT under 2 nearly excludes delirium.
- Teach suicide inquiry as a skill practised aloud until it is comfortable, and teach that its purpose is to hear the answer and act on it, because no scale and no algorithm will do the predicting.
- Teach capacity as a four-part, decision-specific assessment with an instrument, and tell students the number: six in ten incapacitated patients are missed by the clinicians treating them.
- Teach the standardised examinations for catatonia, dyskinesia and withdrawal by demonstration; they are the parts of psychiatry that reproduce like the rest of physical diagnosis.
- Teach the physical examination as part of the psychiatric one, and the collateral history as part of the examination.
- Be honest about the kappa of 0.3, and use it: students who know that two experts agree on depression less often than a questionnaire agrees with an interview will describe more carefully, measure more readily and conclude more slowly — which is what the evidence asks of the psychiatric examination.

References

1. Regier DA, Narrow WE, Clarke DE, et al. DSM-5 field trials in the United States and Canada, part II: test–retest reliability of selected categorical diagnoses. *Am J Psychiatry* 2013;170:59–70.
2. Freedman R, Lewis DA, Michels R, et al. The initial field trials of DSM-5: new blooms and old thorns. *Am J Psychiatry* 2013;170:1–5.

3. Bush G, Fink M, Petrides G, Dowling F, Francis A. Catatonia. I. Rating scale and standardized examination. *Acta Psychiatr Scand* 1996;93:129–36.
4. Assessment of catatonia and inter-rater reliability of three instruments: a descriptive study. *Int J Ment Health Syst* 2021. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8607401/>
5. Sullivan JT, Sykora K, Schneiderman J, Naranjo CA, Sellers EM. Assessment of alcohol withdrawal: the revised Clinical Institute Withdrawal Assessment for Alcohol scale (CIWA-Ar). *Br J Addict* 1989;84:1353–7; and later reliability studies of the scale.
6. Fujikawa M, et al. Diagnosis and symptom assessment in telepsychiatry vs. face-to-face settings: a systematic review and meta-analysis. *Psychiatry Clin Neurosci* 2025. <https://onlinelibrary.wiley.com/doi/10.1111/pcn.13860>
7. Sessums LL, Zembruska H, Jackson JL. Does this patient have medical decision-making capacity? *JAMA* 2011;306:420–7.
8. Benchmarking large language models against practicing clinicians on psychopathological assessment. *NPJ Digit Med* 2026. <https://www.nature.com/articles/s41746-026-02852-7>
9. Williams JW, Noël PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? *JAMA* 2002;287:1160–70.
10. Levis B, Benedetti A, Thombs BD; DEPRESSD Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* 2019;365:l1476.
11. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic meta-analysis. *Gen Hosp Psychiatry* 2016;39:24–31.
12. Kitchens JM. Does this patient have an alcohol problem? *JAMA* 1994;272:1782–7.
13. Buchsbaum DG, Buchanan RG, Centor RM, Schnoll SH, Lawton MJ. Screening for alcohol abuse using CAGE scores and likelihood ratios. *Ann Intern Med* 1991;115:774–7.
14. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Arch Intern Med* 1998;158:1789–95.
15. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *Br J Psychiatry* 2017;210:387–95.
16. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull* 2017;143:187–232.
17. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry* 2019;76:642–51.
18. Machine learning algorithms and their predictive accuracy for suicide and self-harm: systematic review and meta-analysis. 2025. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12425223/>
19. Predictive performance of machine learning for suicide in adolescents: systematic review and meta-analysis. *J Med Internet Res* 2025;27:e73052.
20. Dazzi T, Gribble R, Wessely S, Fear NT. Does asking about suicide and related behaviours induce suicidal ideation? What is the evidence? *Psychol Med* 2014;44:3361–3.
21. Fazel S, Singh JP, Doll H, Grann M. Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: systematic review and meta-analysis. *BMJ* 2012;345:e4692.
22. Performance of automatic speech analysis in detecting depression: systematic review and meta-analysis. *JMIR Ment Health* 2025;12:e67802.
23. Language-based detection of depression with machine learning: systematic review and meta-analysis. *NPJ Digit Med* 2026. <https://www.nature.com/articles/s41746-026-02448-1>
24. The performance of wearable device-based artificial intelligence in detecting depression: systematic review and meta-analysis. *JMIR Ment Health* 2026;13:e85319.
25. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. *Neurosci Biobehav Rev* 2024. <https://www.sciencedirect.com/science/article/pii/S0149763424000095>
26. More than a biomarker: could language be a biosocial marker of psychosis? *Schizophrenia (Heidelb)* 2021. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8408150/>

27. Assessing the accuracy and reliability of large language models in psychiatry using standardized multiple-choice questions: cross-sectional study. *J Med Internet Res* 2025;27:e69910.
28. Diagnostic accuracy of large language models in psychiatry. *Asian J Psychiatr* 2024.
<https://pubmed.ncbi.nlm.nih.gov/39111087/>
29. Large language models for psychiatric diagnosis based on multicenter real-world clinical records: comparative study. *JMIR Med Inform* 2026;14:e77699.

Caveats

The DSM-5 kappa values are those reported for the field trials as summarised in the trial papers and their editorial; the exact figure for major depressive disorder is quoted as about 0.3. The CAGE row is qualitative because the score-by-score likelihood ratios from Buchsbaum's paper were not retrieved. The violence-risk row summarises Fazel's conclusion without its predictive values. The telepsychiatry kappas come from two different pooled analyses within the same review and are reported as such. The wearable and language-model figures are from very recent meta-analyses and benchmarks and will move. Where a reference is given by title and address only, the search results did not return an author list; check before distribution.