

THE ECOSYSTEM OF EVIDENCE

Navigating the future
of the ecosystem of evidence

Artificial intelligence and digital transformation
in research and healthcare

11th EBHC International Conference
Taormina, 21st – 24th October 2026

#EBHC2026

Explainable Artificial Intelligence (XAI) in Mental Disorder Diagnosis

A systematic mapping review

Xi Jia
University College London

30th 1996-2026 GIMBE EVIDENCE FOR HEALTH

International Society for Evidence-Based Health Care

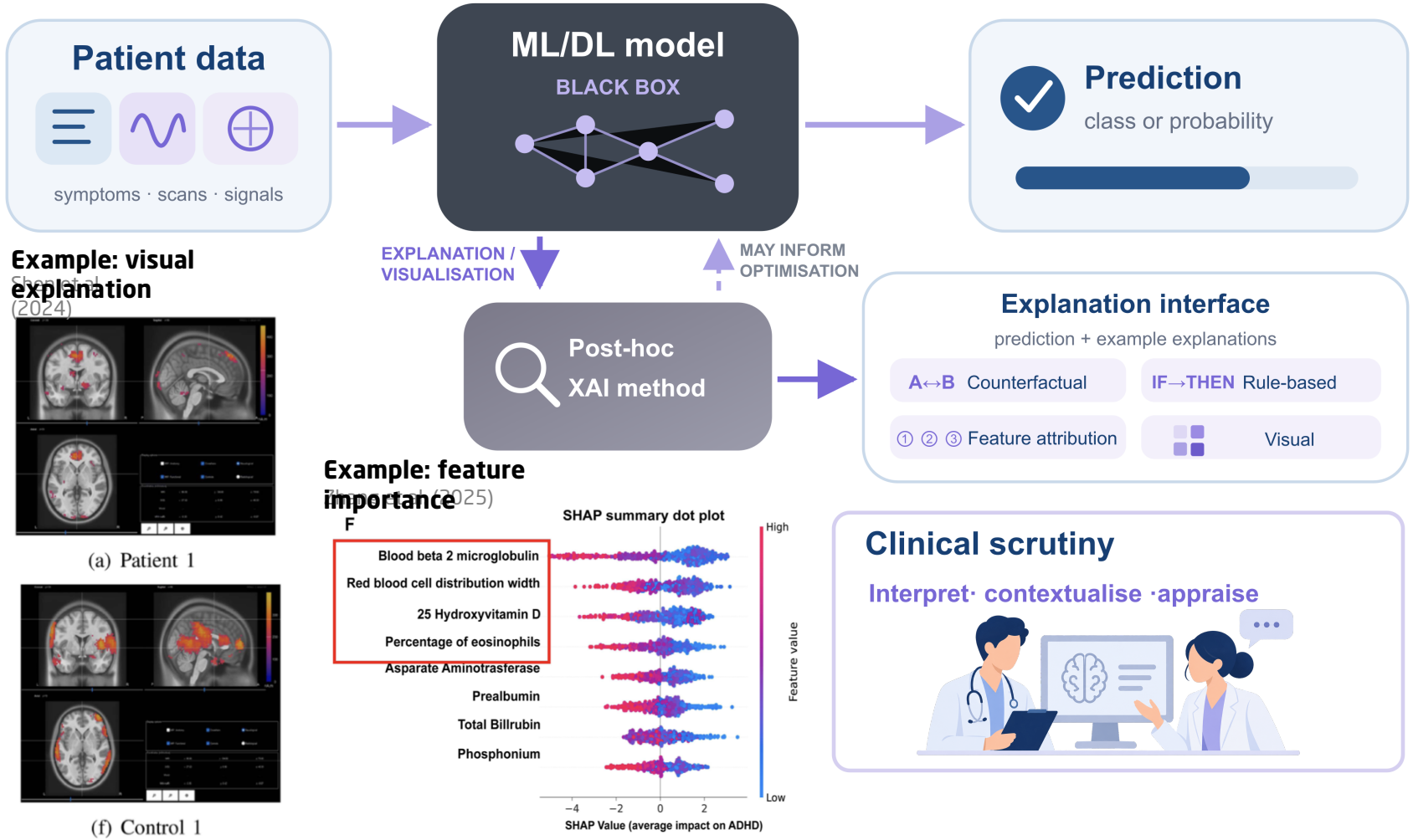
Background

Why this context?

Mental disorder diagnosis often involves:

1. **Heterogeneous presentations**
   Different profiles within the same diagnosis
2. **Psychiatric comorbidity**
   Lower inter-rater and test-retest reliability
3. **Limited objective anchors**
   Few validated biomarkers in routine use

→ **DIAGNOSTIC UNCERTAINTY**



Aims

To systematically map the applications of XAI in mental disorder ~~diag~~nosis

01 **EVIDENCE LANDSCAPE**

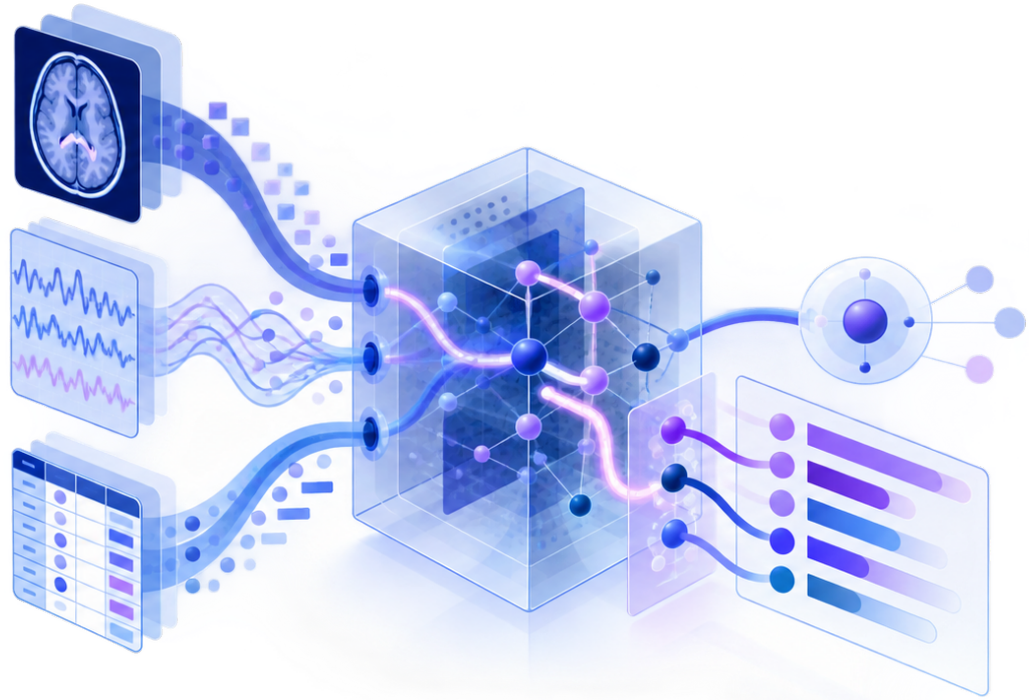
Across which mental disorders and data modalities have XAI techniques been applied in the current evidence base?

02 **ROLE AND OPERATION**

What primary roles does XAI play in supporting mental disorder diagnosis, and how are these roles operationalised through different techniques, scopes and evaluation strategies?

03 **CLINICAL TRANSLATION**

How ready are current XAI-enabled diagnostic models for real-world clinical use, and what limitations or future directions have been identified to support this transition?



Methods

INFORMATION SOURCES

1 Jan 2015–3 Jun 2025

Scopus	PubMed	Web of Science Core Collection	IEEE Xplore	ACM Digital Library

STUDY SELECTION

1,083 records identified → **694** records screened → **246** full texts assessed → **90** studies included

389 duplicates removed

432 excluded · 16 not retrieved

156 full-text reports excluded

ELIGIBILITY

- Post-hoc XAI technique(s)
- Interpret AI models for mental disorder diagnosis
- Be based on clinical data

DATA EXTRACTED

Study characteristics
(e.g., disorder type, data modality)

Methodological features
(e.g., AI/XAI techniques applied, explanatory scope)

Outcomes relevant to the research questions
(e.g., XAI role, XAI evaluation, reported limitations and future directions)

ADDITIONAL ANALYSES

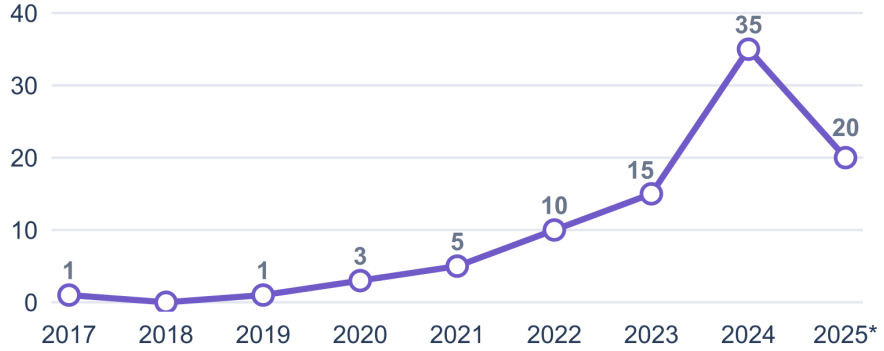
Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)
Risk of bias and Applicability

Collocation Analysis
Reported limitations and future directions

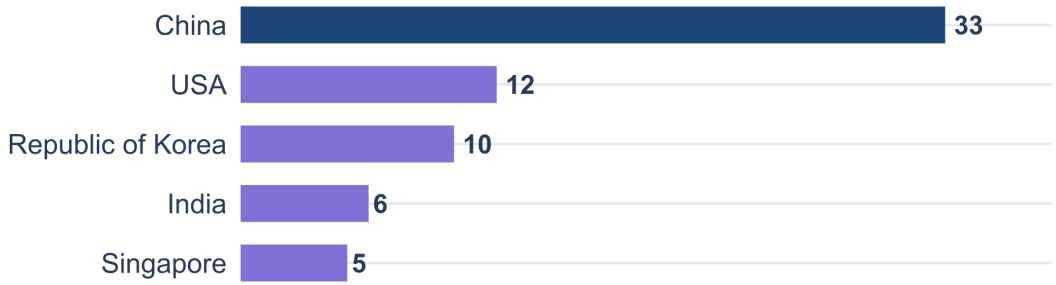
Results

Studies by publication year

*2025 includes studies published up to 3 June



Studies by country (top five)



Mental disorder × XAI technique

	SHAP	Grad-CAM	LIME	LRP	Saliency
Anxiety disorders	5		1		
Bipolar disorder	5	1	1		
Depression	23	3	3	2	
Disruptive behaviour disorders		3			
Eating disorders					
Neurodevelopmental disorders	18	4	2	1	1
PTSD	1				
Schizophrenia	11	4	3	6	3

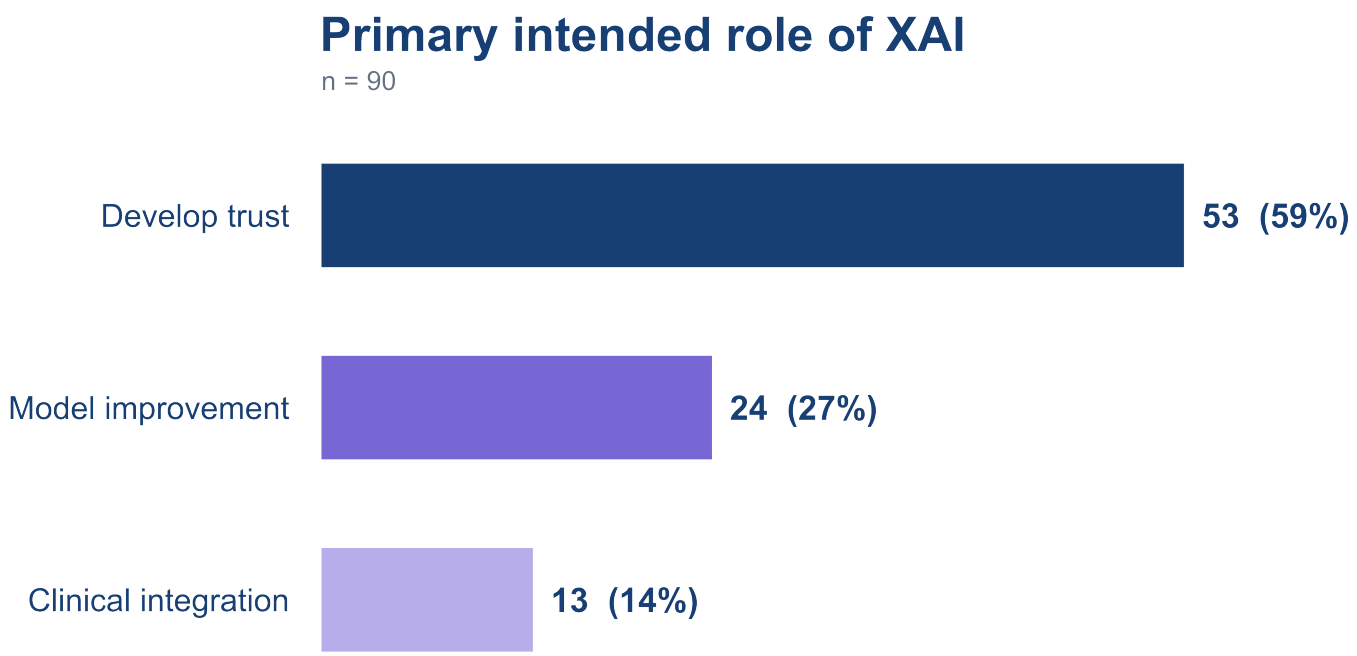
Data modality × XAI technique

	SHAP	Grad-CAM	LIME	LRP	Saliency
Behavioural data	3	1			
Clinical attributes	11		3		
Multimodal data	11	5		1	
Neuroimaging data	16	1	1	4	3
(Neuro)physiological signals	12	5	3	3	1
Omics data	1		2		
Textual data	1		1		

Heatmaps show the five most frequently used XAI techniques; counts are technique occurrences, and one study may contribute more than once.

Results

XAI was used mainly to "develop trust", but trust was rarely demonstrated



Role classification adapted from Ali et al. (2023)

THE EVALUATION GAP

Only 23 / 90 included studies evaluated XAI output

DEVLOP TRUST	**10/53**	1 combined human-centred feedback with computational checks
MODEL IMPROVEMENT	**9/24**	All computer-centred assessments (e.g., fidelity, stability)
CLINICAL INTEGRATION	**4/13**	1 human-centred assessment (e.g., perceived usability)

Results

BENCH **Most XAI studies remain at the "bench"** BEDSIDE

82/90
Internal validation
Performance assessed within the original development data

22 biomarker-focused studies (9 discovery · 13 validation)

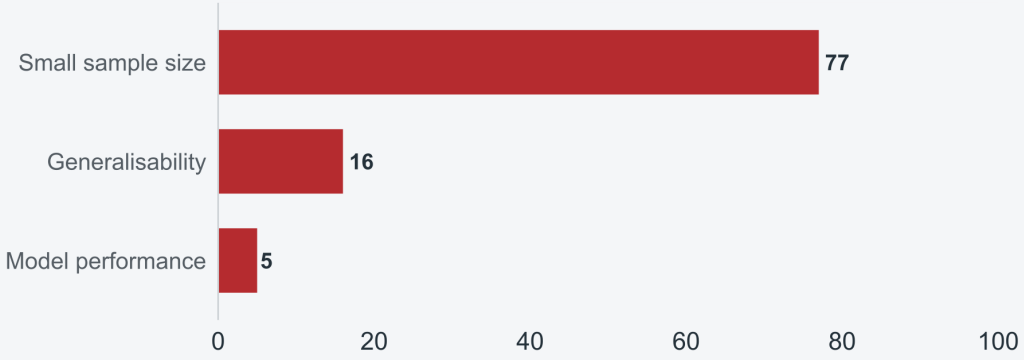
7/90
External validation
Performance tested using fully independent data

2 decision-support tools both focused on ADHD

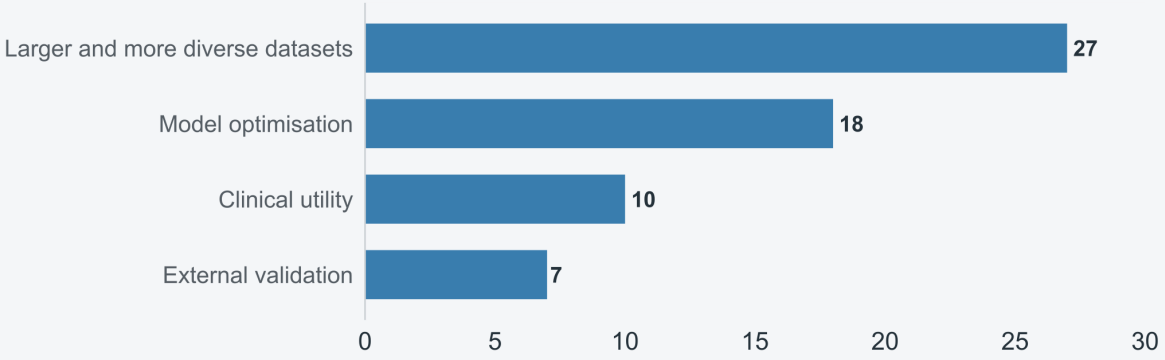
1/90
Real-world evaluation
Prospective testing in a real-world setting using real-time data

Depression prototype greater trust · poorer usability

Most frequently reported limitations



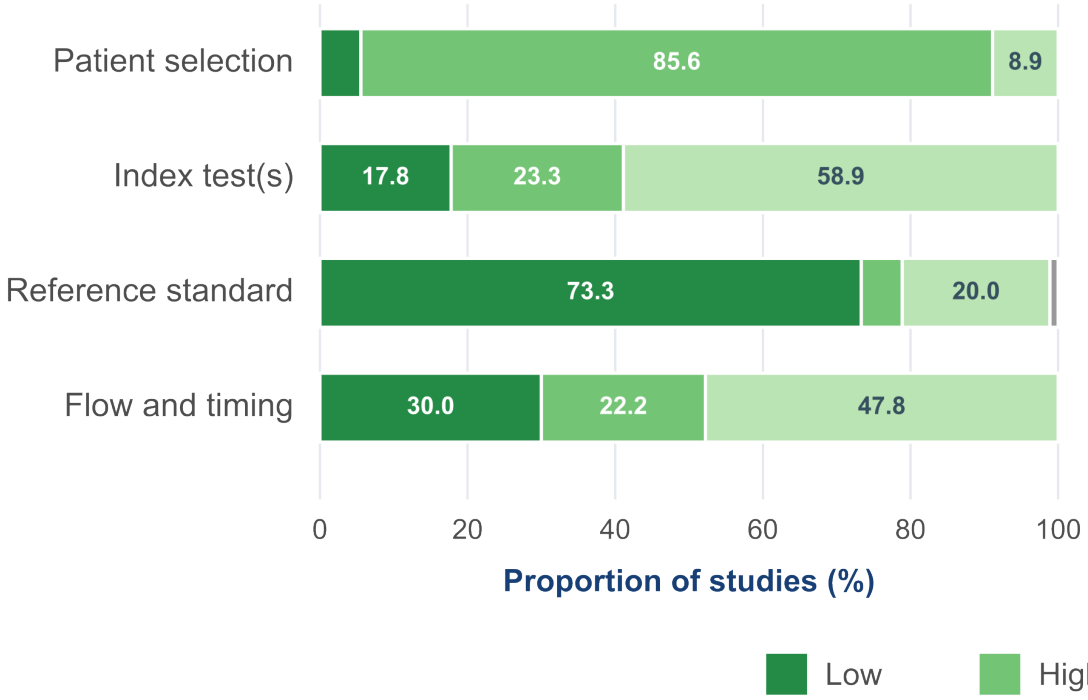
Most frequently proposed directions



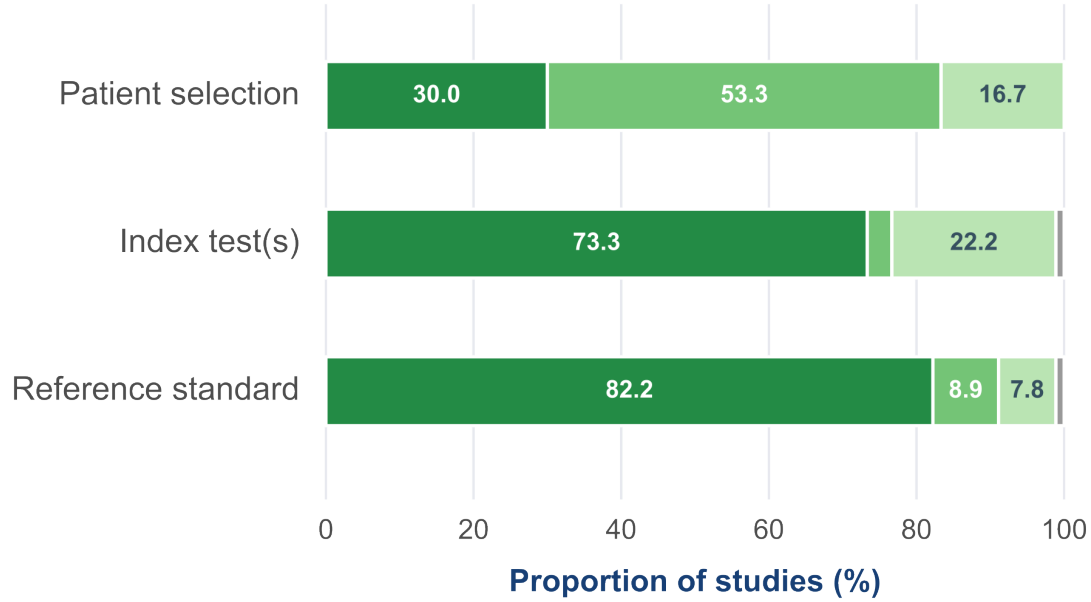
Results

Patient selection was the clearest concern, with substantial uncertainty around index test(s)

Risk of bias



Applicability



Limits

- **Search coverage**

  Five databases were searched, but no supplementary or grey literature searching was undertaken

- **Single-reviewer**

  This may have increased the potential for subjective interpretation or omission

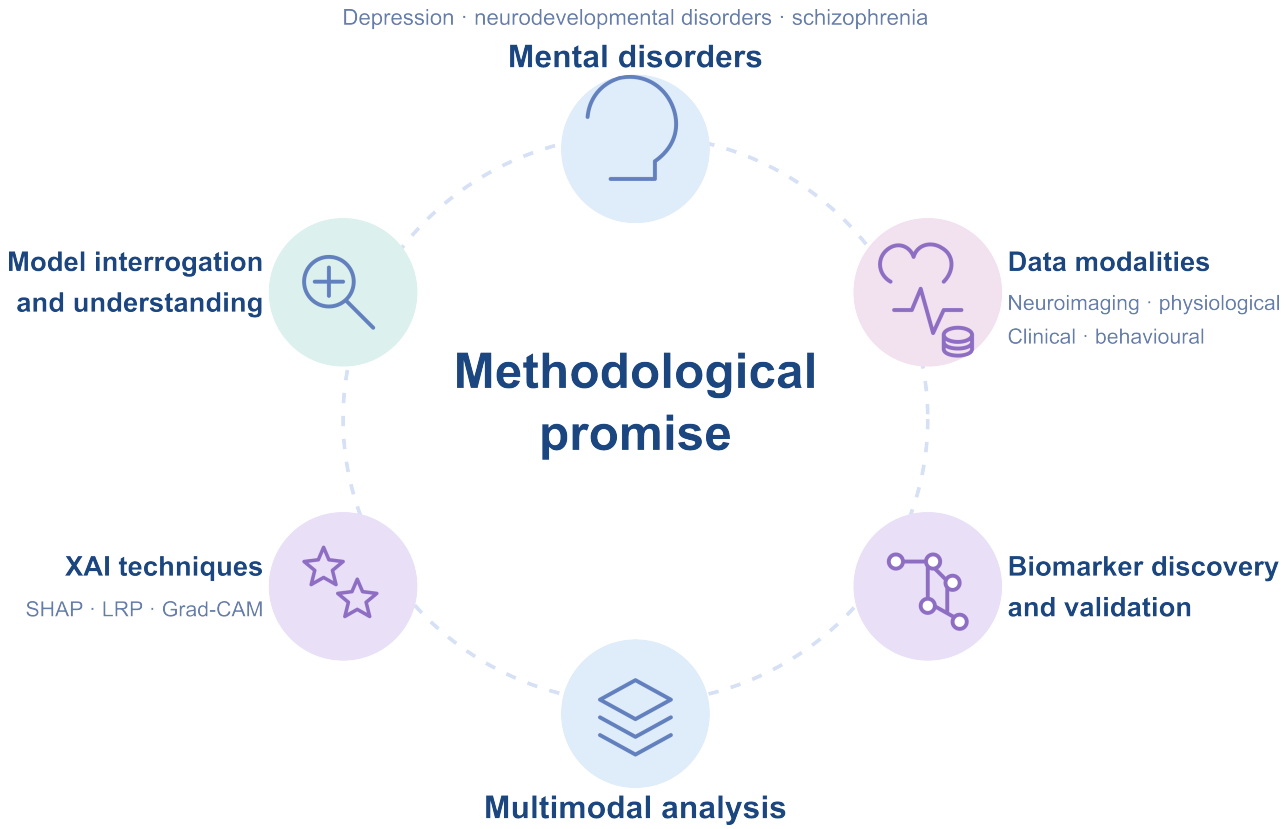
- **Predefined framework**

  The three XAI roles may not have captured every function reported in the literature

- **Language restriction**

  English-only inclusion may have excluded relevant evident

Conclusions



FUTURE IMPLICATIONS

01 Evaluate explanations

Technical soundness · clinical relevance

E.g. CLIX-M, intended-user evaluation

02 Build evidence for clinical use

External validation · real-world evaluation

E.g. silent trials, clinician and patient involvement

03 Report transparently

Patient selection · index-test procedures

E.g. STARD-AI, DECIDE-AI

BROADER IMPLICATION FOR EVIDENCE SYNTHESIS

Make underlying assumptions visible

What counts as working?

For whom?

What is left out?