Supplementary methods. Search strategy and scope of the Review

This file documents the literature search supporting the critical narrative Review Relapse prediction after remission in major depressive disorder: From single classification to sequential clinical decisions. The search was designed primarily to characterise artificial-intelligence (AI)-based prediction of relapse or recurrence after remission in major depressive disorder (MDD). Prognostic modelling, digital phenotyping, neuroimaging, electronic health record (EHR) data, and speech or language markers were searched secondarily, to establish what each data stream can credibly support. It was not designed or reported as a PRISMA systematic review or meta-analysis: no pooled effect estimates were computed, and the search was not intended as an exhaustive inventory of all prognostic studies in depression, but to define the scope of the evidence base, identify the data modalities recurring in this literature, and support a critical mapping between evidence streams and the Stratify-Monitor-Confirm (SMC) framework.

Searches were run in PubMed and Web of Science on 23 April 2026, restricted to English-language records. A primary search combined terms for MDD, the remitted or maintenance state, relapse-related outcomes (relapse, recurrence, readmission, retreatment, future episode), and prediction or AI/machine-learning methods. Because direct prospective evidence on relapse after remission is limited, modality-specific supplementary searches were added for EHR and readmission studies, digital phenotyping and passive sensing, speech/voice/text/NLP, and prognostic neuroimaging. Adjacent evidence (cross-sectional depressive state, illness course, future depressive episode, treatment re-entry, or symptom trajectory) was admitted only where it informed the clinical question a given modality could credibly answer; these endpoints were not treated as equivalent to relapse after remission. The formal literature-search cut-off was 23 April 2026. The publication status of the Puder *et al* 2025 report was rechecked on 17 September 2026. The full report remained a medRxiv preprint; a related conference abstract was identified, but no peer-reviewed full-length journal article was found.

Records from the two databases were combined and de-duplicated by DOI and title, yielding 2,964 unique records. Title-and-abstract screening excluded 2,239 records, leaving 725 potentially relevant reports for full-text triage or targeted checking. Screening was conducted by Chen SY without independent duplicate screening, within a critical narrative-review framework rather than as a formal systematic review; uncertain or borderline reports were rechecked at full text against the stated scope criteria during revision. Co-authors contributed to critical interpretation of the synthesis during manuscript revision but did not repeat the screening process. References identified during targeted checking were considered against the same scope criteria; no formal backward- or forward-citation-tracking procedure was undertaken. The only preprint retained was Puder *et al*,

because it directly addressed the review question and no peer-reviewed full-length journal article was identified; it is identified as a preprint wherever used. Conference abstracts and other unpublished reports were not retained in the synthesis. The related Puder conference abstract was checked for publication status but was not retained as a separate report. Because study selection and citation were conducted within a critical narrative-review framework, selective-citation bias cannot be excluded. Exclusion at the title-and-abstract stage was based on prespecified scope criteria: not MDD or depressive disorder; not post-remission relapse, recurrence, readmission, retreatment, future episode, course, monitoring, or directly relevant adjacent evidence; not a predictive, prognostic, artificial-intelligence, machine-learning, or digital-biomarker study; acute treatment response only; non-human or basic-science evidence; non-original article except key reviews, definitions, or methodological papers retained as contextual evidence; non-English record; duplicate; or unusable record. Nine primary reports formed the core AI/ML synthesis: original reports in which an individual-level AI/ML or statistical-learning prediction model was developed or validated in an MDD-specific analytic cohort for a future relapse-relevant outcome. Core membership required the model-development or validation population to be MDD-specific; MDD subgroup results from models developed in mixed mood-disorder cohorts were treated as supporting evidence. Reports retained outside that core were used to interpret the clinical role, temporal characteristics, and limitations of individual data streams; 27 study reports in total are mapped in Supplementary Table 4. For synthesis, reports were not treated as formal evidence grades. Core membership and evidence directness were coded separately. Evidence directness reflects the endpoint addressed by the report rather than whether the study belongs to the core AI/ML set. Direct relapse-relevant evidence comprised reports in which the outcome was a prospectively or clinically observed relapse, recurrence, readmission, or retreatment event relevant to the review target. Population specificity was recorded separately and did not by itself determine endpoint directness. Adjacent modality-specific evidence comprised reports informing one Stratify-Monitor-Confirm layer but using a non-equivalent outcome such as remission status, symptom state or trajectory, first onset, or concurrent diagnosis rather than an observed relapse-relevant event. Contextual/methodological evidence comprised consensus definitions, reviews, methodological papers, feasibility and implementation studies, and background evidence used to frame terminology, model evaluation, implementation, or clinical risk. Consistent with a narrative review, appraisal was qualitative and structured around the dimensions reported in Supplementary Table 4: population, sample and design; outcome definition, events and follow-up; validation design; reported performance where available; and the clinical question each report could credibly support. No formal risk-of-bias scoring or quantitative synthesis was undertaken.

Supplementary Table 1 Search blocks and records retrieved

Concept block	PubMed (n)	Web of Science (n)
Core relapse/recurrence prediction after remission (AI/ML and prognostic methods)	525	173
Electronic health records, claims, and readmission	95	362
Digital phenotyping, wearable, actigraphy, ecological momentary assessment	368	724
Speech, voice, language, natural language processing	319	527
Neuroimaging (prognostic)	265	426

Note: Figures are per-search, pre-deduplication retrieval counts. PubMed block-specific counts were reproduced on 26 June 2026 and may differ slightly from the original 23 April 2026 search because of database updates; this count check did not constitute a search update. Web of Science counts and de-duplicated totals reflect the original search log. After combining both databases and de-duplicating by DOI and title, 2,964 unique records remained. Full search strings are given in Supplementary Table 2. AI: artificial intelligence; EHR: electronic health record; EMA: ecological momentary assessment; ML: machine learning; NLP: natural language processing.

Supplementary Table 2 Full electronic search strategies

Search	Search strategy
PubMed: Main	("Major Depressive Disorder"[MeSH Terms] OR "major depressive disorder"[tiab] OR "major depression"[tiab] OR "unipolar depression"[tiab] OR "unipolar depressive"[tiab] OR MDD[tiab]) AND (remission[tiab] OR remitted[tiab] OR recovered[tiab] OR recovery[tiab] OR euthymic[tiab] OR "maintenance phase"[tiab] OR "maintenance treatment"[tiab] OR "continuation phase"[tiab] OR "continuation treatment"[tiab] OR "post-remission"[tiab] OR "post remission"[tiab]) AND ("Recurrence"[MeSH Terms] OR relapse[tiab] OR relapses[tiab] OR relapsing[tiab] OR recurrence[tiab] OR recurrences[tiab] OR recurrent[tiab] OR "symptom return"[tiab] OR "symptom re-emergence"[tiab] OR "symptom reemergence"[tiab] OR "symptom recurrence"[tiab] OR readmission[tiab] OR readmissions[tiab] OR rehospitalization[tiab] OR rehospitalisation[tiab] OR retreatment[tiab] OR "re-treatment"[tiab] OR "future episode"[tiab] OR "future episodes"[tiab] OR "subsequent episode"[tiab] OR "subsequent episodes"[tiab] OR "new episode"[tiab] OR "new episodes"[tiab] OR "time to relapse"[tiab] OR "time-to-relapse"[tiab] OR "time to recurrence"[tiab] OR "time-to-recurrence"[tiab]) AND ("Machine Learning"[MeSH Terms] OR "Artificial Intelligence"[MeSH Terms] OR "Deep Learning"[MeSH Terms] OR "machine learning"[tiab] OR "machine-learning"[tiab] OR "deep learning"[tiab] OR "artificial intelligence"[tiab] OR "neural network"[tiab] OR "neural networks"[tiab] OR "random forest"[tiab] OR "random forests"[tiab] OR "support vector machine"[tiab] OR "gradient boosting"[tiab] OR XGBoost[tiab] OR LASSO[tiab] OR "elastic net"[tiab] OR "ensemble model"[tiab] OR "ensemble learning"[tiab] OR "prediction model"[tiab] OR "predictive model"[tiab] OR "prognostic model"[tiab] OR "risk prediction"[tiab] OR "risk stratification"[tiab] OR predict[tiab] OR predicts[tiab] OR predicted[tiab] OR predicting[tiab] OR predictive[tiab] OR prediction[tiab] OR prognostic[tiab] OR prognosis[tiab]) AND English[Language]

Search	Search strategy
PubMed: Supplementary A (EHR / readmission)	("Major Depressive Disorder"[MeSH Terms] OR "major depressive disorder"[tiab] OR "major depression"[tiab] OR MDD[tiab] OR "depressive disorder"[tiab]) AND (readmission[tiab] OR readmissions[tiab] OR rehospitalization[tiab] OR rehospitalisation[tiab] OR "re-hospitalization"[tiab] OR "re-hospitalisation"[tiab] OR retreatment[tiab] OR "re-treatment"[tiab] OR "psychiatric readmission"[tiab] OR "hospital readmission"[tiab] OR "30-day readmission"[tiab] OR "post-discharge"[tiab] OR "post discharge"[tiab]) AND (predict[tiab] OR prediction[tiab] OR predictive[tiab] OR prognostic[tiab] OR prognosis[tiab] OR "risk prediction"[tiab] OR "risk factor"[tiab] OR "risk factors"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "artificial intelligence"[tiab] OR "natural language processing"[tiab] OR NLP[tiab] OR "prediction model"[tiab] OR "prognostic model"[tiab]) AND English[Language]
PubMed: Supplementary B (digital phenotyping / wearable / EMA)	("Major Depressive Disorder"[MeSH Terms] OR "major depressive disorder"[tiab] OR "major depression"[tiab] OR MDD[tiab] OR "mood disorder"[tiab] OR "mood disorders"[tiab]) AND ("digital phenotyping"[tiab] OR "passive sensing"[tiab] OR "mobile sensing"[tiab] OR smartphone[tiab] OR smartphones[tiab] OR "mobile phone"[tiab] OR wearable[tiab] OR wearables[tiab] OR actigraphy[tiab] OR actigraph[tiab] OR accelerometer[tiab] OR "ecological momentary assessment"[tiab] OR EMA[tiab] OR "experience sampling"[tiab] OR GPS[tiab] OR "sensor data"[tiab]) AND (relapse[tiab] OR recurrence[tiab] OR recurrent[tiab] OR "mood episode"[tiab] OR "mood episodes"[tiab] OR "symptom trajectory"[tiab] OR "symptom trajectories"[tiab] OR "early warning"[tiab] OR "critical slowing"[tiab] OR "depressive symptoms"[tiab] OR "symptom severity"[tiab] OR "mood prediction"[tiab] OR "mood state"[tiab] OR "mood states"[tiab]) AND English[Language]

Search	Search strategy
PubMed: Supplementary C (speech / voice / text / NLP)	("Major Depressive Disorder"[MeSH Terms] OR "major depressive disorder"[tiab] OR "major depression"[tiab] OR MDD[tiab] OR "depressive disorder"[tiab] OR "depressive episode"[tiab] OR "depressive episodes"[tiab]) AND (speech[tiab] OR voice[tiab] OR vocal[tiab] OR acoustic[tiab] OR prosody[tiab] OR prosodic[tiab] OR "linguistic marker"[tiab] OR "linguistic markers"[tiab] OR "linguistic feature"[tiab] OR "linguistic features"[tiab] OR "social media"[tiab] OR "text analysis"[tiab] OR "natural language processing"[tiab] OR NLP[tiab]) AND (predict[tiab] OR prediction[tiab] OR predictive[tiab] OR detect[tiab] OR detection[tiab] OR severity[tiab] OR "clinical state"[tiab] OR longitudinal[tiab] OR prospective[tiab] OR classification[tiab] OR classifier[tiab]) AND English[Language]
PubMed: Supplementary D (neuroimaging)	("Major Depressive Disorder"[MeSH Terms] OR "major depressive disorder"[tiab] OR "major depression"[tiab] OR MDD[tiab]) AND (remission[tiab] OR remitted[tiab] OR recovered[tiab] OR recovery[tiab] OR euthymic[tiab] OR "follow-up"[tiab] OR "follow up"[tiab] OR longitudinal[tiab] OR prospective[tiab]) AND (relapse[tiab] OR relapses[tiab] OR recurrence[tiab] OR recurrences[tiab] OR recurrent[tiab] OR "recurring episode"[tiab] OR "recurring episodes"[tiab] OR "future recurrence"[tiab] OR "future episode"[tiab] OR "future episodes"[tiab] OR "future depressive episode"[tiab] OR "future depressive episodes"[tiab] OR resilience[tiab] OR "vulnerability to recurrence"[tiab] OR "risk of recurrence"[tiab] OR "risk of relapse"[tiab]) AND ("functional magnetic resonance imaging"[tiab] OR fMRI[tiab] OR "resting state"[tiab] OR "resting-state"[tiab] OR "functional connectivity"[tiab] OR "structural MRI"[tiab] OR neuroimaging[tiab] OR "neural signature"[tiab] OR "brain connectivity"[tiab] OR "cortical thickness"[tiab] OR "gray matter"[tiab] OR "grey matter"[tiab] OR "white matter"[tiab] OR EEG[tiab] OR electroencephalograph*[tiab] OR "diffusion tensor"[tiab] OR DTI[tiab] OR PET[tiab] OR "positron emission"[tiab]) AND English[Language]

Search	Search strategy
Web of Science: Main	TS=(("major depressive disorder" OR "major depression" OR "unipolar depression" OR "unipolar depressive" OR "MDD" OR "depressive disorder" OR "depressive episode") AND (remission OR remitted OR recovered OR recovery OR euthymic OR "maintenance phase" OR "maintenance treatment" OR "continuation phase" OR "continuation treatment" OR "post-remission") AND (relaps* OR recurrenc* OR recurrent OR "symptom return" OR "symptom re-emergence" OR "symptom recurrence" OR readmission* OR rehospitali* OR retreatment OR "re-treatment" OR "future episode" OR "subsequent episode" OR "new episode" OR "time to relapse" OR "time to recurrence") AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network" OR "neural networks" OR "random forest" OR "support vector machine" OR "gradient boosting" OR "XGBoost" OR "LASSO" OR "elastic net" OR "ensemble model" OR "ensemble learning" OR "prediction model" OR "predictive model" OR "prognostic model" OR "risk prediction" OR "risk stratification" OR predict* OR prognostic OR prognosis)) AND LA=(English) This main strategy was superseded by the five block searches (A–E); retrieval counts are reported for those blocks in Supplementary Table 1.
Web of Science: Search A (core relapse prediction; n = 173)	("major depressive disorder" OR "major depression" OR "unipolar depression" OR ("MDD" AND depress*)) AND (relapse OR recurrence OR recurrent OR readmission OR rehospitali* OR "future episode*" OR "illness course" OR "disease course" OR "post-remission" OR "maintenance phase") AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "random forest" OR "support vector" OR "neural network*" OR "gradient boosting" OR "prediction model*" OR "prognostic model*" OR "risk prediction" OR "risk stratification" OR "digital biomarker*" OR "computational psychiatry")

Search	Search strategy
Web of Science: Search B (digital phenotyping / wearable / EMA; n = 724)	("major depressive disorder" OR "major depression" OR "unipolar depression" OR "depressive disorder*" OR "depressive episode*" OR ("MDD" AND depress*)) AND ("digital phenotyping" OR smartphone* OR "mobile phone*" OR wearable* OR actigraph* OR acceleromet* OR "passive sensing" OR "passive monitor*" OR "ecological momentary assessment" OR EMA OR "experience sampling" OR "rest-activity" OR "circadian rhythm*" OR "sleep-wake") AND (predict* OR trajectory OR longitudinal OR "within-person" OR "individual-level" OR idiographic OR personalized OR personalised OR relapse OR recurrence OR "symptom change" OR "mood episode*" OR "early warning" OR "critical slowing")
Web of Science: Search C (speech / voice / text / NLP; n = 527)	("major depressive disorder" OR "major depression" OR "unipolar depression" OR "depressive disorder*" OR "depressive episode*" OR ("MDD" AND depress*)) AND (speech OR "voice analysis" OR vocal OR acoustic* OR prosod* OR "natural language processing" OR NLP OR "text analysis" OR "language model*" OR "linguistic marker*" OR "sentiment analysis" OR "social media" OR "written language") AND (severity OR "clinical state" OR symptom* OR longitudinal OR "state-sensitive" OR biomarker* OR relapse OR recurrence OR "mood change*" OR "symptom change")
Web of Science: Search D (neuroimaging prognostic; n = 426)	("major depressive disorder" OR "major depression" OR "unipolar depression" OR ("MDD" AND depress*)) AND (fMRI OR "functional MRI" OR "resting-state" OR "functional connectivity" OR "structural MRI" OR "cortical thickness" OR "brain volume" OR DTI OR neuroimaging OR "brain imaging") AND (relapse OR recurrence OR recurrent OR "future episode*" OR "post-remission" OR remitted OR remission) AND (predict* OR prognos* OR course OR resilience OR vulnerabilit*)

Search	Search strategy
Web of Science: Search E (EHR / claims / readmission; n = 362)	("major depressive disorder" OR "major depression" OR "unipolar depression" OR ("MDD" AND depress*)) AND ("electronic health record*" OR EHR OR "medical record*" OR "claims data" OR "administrative data" OR "clinical record*" OR "clinical note*" OR "discharge summar*" OR "natural language processing" OR NLP) AND (relapse OR recurrence OR readmission OR rehospitali* OR "psychiatric readmission" OR retreatment OR "illness course" OR "disease course" OR predict* OR prognos* OR "risk model*")

Note: Search strings are reproduced verbatim from the electronic search log. Database-specific syntax was retained. AI: artificial intelligence; EHR: electronic health record; EMA: ecological momentary assessment; MDD: major depressive disorder; NLP: natural language processing.

Supplementary Table 3 Endpoint heterogeneity across the reviewed evidence and its implications for Stratify-Monitor-Confirm evaluation

Endpoint type	Typical operationalisation	What it captures	What it does not capture	Implication for SMC evaluation
DSM-defined relapse or recurrence	Structured interview or clinical adjudication during prospective follow-up	Return to full syndromal depression	Subsyndromal drift and between-visit change preceding the event	Most appropriate as the downstream outcome against which the Confirm layer, and the sequence as a whole, are judged
Psychiatric readmission	Inpatient admission recorded in electronic health records or administrative data	Severe deterioration that reaches inpatient care	Relapse managed in outpatient or primary care; admission also reflects help-seeking and service access	Informative for Stratify (baseline risk), but a partial and access-dependent proxy for relapse
Retreatment or treatment re-entry	Claims, prescription records, or return to depression care	Renewed clinical contact or perceived treatment need	Whether symptoms actually met relapse criteria at re-entry	Useful for scalable baseline stratification, not for confirming a clinical transition
Symptom-course or remission status	Rating-scale trajectory, or remission versus non-remission at a fixed follow-up window	Longitudinal symptom burden and course	The timing of, and transition into, a discrete relapse event	Adjacent evidence for baseline vulnerability; calibrates Stratify but is not equivalent to relapse

Endpoint type	Typical operationalisation	What it captures	What it does not capture	Implication for SMC evaluation
Cross-sectional state or diagnosis	Case versus control, concurrent symptom severity, or speech/text state markers at a single time point	Sensitivity to current depressive state under defined conditions	Any prospective transition from remission to relapse	Candidate evidence for a triggered Confirm probe, not for relapse prediction in itself
Constructed relapse or recurrence	Algorithmic threshold or distance rule applied to cross-sectional or proxy data	Feasibility of modelling transition-like patterns	A real, prospectively observed clinical transition	Methodological and adjacent evidence only; cannot stand in for prospective relapse evidence at any layer

Note: This table distinguishes non-equivalent outcome constructs and indicates their relevance to evaluation of the proposed SMC stages. Stage-specific evaluation should align each clinical decision with a prespecified endpoint, while evaluation of the full sequence requires a consistently assessed clinical relapse outcome. This table classifies endpoint types, not individual studies; the study-level mapping is provided in Supplementary Table 4. DSM: Diagnostic and Statistical Manual of Mental Disorders; SMC: Stratify-Monitor-Confirm.

Supplementary Table 4 Detailed evidence map of core AI/ML relapse-relevant reports and supporting evidence

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Lythe[19]; Lawrence[20]	Core AI/ML: Direct relapse-relevant	Same prospective UK cohort of medication-free, fully remitted MDD without relevant comorbidity; self-blame task fMRI. Lythe: 75 followed; primary imaging comparison n=56 (31 stable, 25 recurrent). Lawrence complete-case model n=52 (34 stable, 18 recurrent).	DSM-IV major depressive episode (MDE) recurrence assessed with blinded LIFE interviews at 3, 6 and 14 months.	Lythe: predictive LDA with LOOCV after same-cohort feature derivation. Lawrence: elastic-net with repeated nested 10-fold CV (inner tuning, outer performance estimation). No external validation (single cohort).	Lythe: cross-validated accuracy 75%, PPV 74%, NPV 76%. Lawrence: clinical-only AUC 0.73; clinical-plus-fMRI AUC 0.90 (accuracy 83%) and all-modalities AUC 0.86 (accuracy 80%). For clinical-plus-fMRI and all-modalities models, respectively, calibration slopes were 0.61/0.55, intercepts 0.02/−0.08, and Brier scores 0.13/0.15. Mechanism-selected signal is promising, but estimates come from one small cohort.	Who remains vulnerable after remission? Development-stage prognostic evidence, not a clinically validated test.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
van Kleef[21]	Core AI/ML: Direct relapse-relevant	Remitted recurrent MDD, medication-free; resting-state fMRI and separately assessed self-referential processing. Baseline n=48 MDD plus 23 controls; prospective relapse analysis n=40 (12 relapse, 28 no relapse).	MDE relapse and time to relapse reconstructed monthly with an 18-month SCID life-chart interview.	Exploratory 70/30 train–test split with SMOTE applied to training data; five classifiers. No external validation (single cohort); very small event and test samples.	Best reported test accuracy 83.3% and F1 0.67 (random forest), with model-dependent accuracy 50–83.3%. Ventral default-mode connectivity extends self-blame findings, but performance is preliminary.	Who remains vulnerable after remission? Requires replication in a larger independent cohort.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Workman[22]; Langenecker[23]; Dunlop[24]	Supporting: Direct relapse-relevant	Prospective fMRI cohorts. Workman: n=47 remitted MDD (17 recurrent, 30 resilient). Langenecker: n=55 with outcome data (21 recurrent, 34 resilient; 60 enrolled). Dunlop: 61 treatment remitters with usable week-12 scans (9 recurrences).	Workman: DSM recurrence by LIFE through 14 months. Langenecker: MDE recurrence by LIFE over 52–72 weeks. Dunlop: quarterly assessment up to 21 months; recurrence after month 9 by LIFE score, sustained HAM-D elevation, treatment change or high suicide risk.	Prospective group-level association analyses; no individual prediction model. No external validation. Langenecker test–retest analyses concern marker reliability, not prognostic validation.	No individual-level discrimination or calibration metrics. Convergence across subgenual, cognitive-control, salience and affective networks is association-level and event counts are small, especially in Dunlop.	Which baseline vulnerability circuits are relapse-relevant? Mechanistic convergence does not establish deployable prediction.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Shu[25]; Dong[26]	Supporting: Direct relapse-relevant	Challenge-task fMRI. Shu: n=42 remitted depression (35 MDD, 7 bipolar depression), 16 relapses (15 MDD, 1 bipolar). Dong: n=74 first-episode, unmedicated MDD at an acute baseline (31 relapse, 43 no relapse).	Shu: study-recorded depressive relapse over 2 years (9 within 6 months); operational diagnostic criteria were not specified in the report. Dong: new DSM-IV-TR MDE after at least 2 months of remission, assessed by two psychiatrists at 2-year telephone follow-up.	Shu: LOOCV used to select the elastic-net penalty, without an independent or nested test cohort; bootstrap CIs. Dong: association/logistic-regression analysis, not a validated prediction model. No external validation.	Shu 2-year model: AUC 0.70 (95% CI 0.54–0.85), accuracy 64.3%, balanced accuracy 59.1%; 6-month AUC 0.61 (0.41–0.80). Dong reported associations, not discrimination or calibration.	Mechanism-aligned evidence, but Shu includes bipolar depression and Dong does not start from remission.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Galioulline[16]	Supporting: Adjacent modality-specific	Balanced UK Biobank sample without prior depression, n=906 (453 with subsequent depression indications, 453 without); resting-state fMRI connectivity.	Future first depressive episode, not post-remission relapse.	Held-out out-of-sample testing. No external validation.	AUC approximately 0.64; a benchmark showing modest discrimination from broad rs-fMRI features.	Modality benchmark only; first onset is not equivalent to recurrence after remission.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Puder[17]	Core AI/ML: Direct relapse-relevant	MACS longitudinal lifetime-MDD cohort, n=346 (240 female); baseline acute n=127, partial remission n=107, full remission n=112. Clinical variables plus 12,090 rs-fMRI connectivity features. medRxiv preprint.	At least one additional MDE in 171/346 participants; none in 175. Mean follow-up 2.18 years (range 1.83–2.99); relapse and recurrence were not separated. Baseline was not uniformly remitted. Psychosocial functioning was a separate outcome.	Random forests with nested CV: 10-fold inner tuning and stratified 10-fold outer evaluation; imputation and feature processing within training folds. No external validation.	Balanced accuracy, mean (SD) across outer folds: reference 57.70% (6.62), clinical 64.64% (10.67), rs-fMRI 51.72% (5.58), multimodal 63.50% (9.91). Clinical versus reference p=0.062; multimodal versus clinical p=0.578. SDs are not CIs. rs-fMRI did not improve discrimination in this sample.	Does broad rs-fMRI add prognostic information? Provisional evidence in a mixed baseline-state MDD cohort; the reported performance is not specific to fully remitted patients. Source locations: pp. 5–8; Table 1, p. 19; Table 2, p. 20; Table 3, p. 21.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Teutenberg[29]	Core AI/ML: Direct relapse-relevant	MACS source sample, n=571 patients with MDD; recurrence analysis n=546 (270 recurrence, 276 no recurrence); trajectory analysis n=557 (discovery 273, validation 284); random-forest models using seven established risk factors and broader exploratory features.	Depressive-episode recurrence and symptom-course trajectories within 2 years.	Recurrence: nested cross-validation with five inner and five outer folds. Trajectories: models trained in the Münster discovery sample and evaluated in the Marburg validation sample within the same two-site study. No independent external cohort.	Seven-factor recurrence model balanced accuracy 62.83%; broader features produced only a minor increase. Validation-sample one-versus-rest trajectory balanced accuracy was 64.23% for remitted and 63.17% for severe courses, but poorer for intermediate trajectories.	How far can established clinical risk factors stratify course? Moderate performance remains below clinical-use requirements.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Zhu[30]; von Stackelberg[31]	Core AI/ML: Direct relapse-relevant	Administrative/EHR cohorts. Zhu: 13,177 patients discharged after an index MDD hospitalisation in one Chinese health system. von Stackelberg: Netherlands administrative source cohort n=106,067; modelling sample 29,019 adults with six monitoring landmarks.	Zhu: any psychiatric readmission at 30, 60, 90, 180 or 365 days after discharge. von Stackelberg: new depression-treatment trajectory after at least a 1-month gap, or suicide, within 12 months; event prevalence 3.75–5.96% across landmarks.	Zhu: temporally separated held-out test sets after inner CV. von Stackelberg: stratified 4-fold internal CV; predictor selection for the simple model used an independent training sample. No external validation (no independent health system).	Zhu random-forest AUCs with 90% CIs from 30 to 365 days: 0.814 (0.758–0.870), 0.780 (0.728–0.833), 0.798 (0.750–0.846), 0.740 (0.687–0.794) and 0.711 (0.676–0.747). von Stackelberg mean AUC: simple logistic 0.653, lasso 0.653, XGBoost 0.649; simple-model mean calibration slope 0.988. Added complexity did not improve prediction.	Scalable risk near a clinical contact, but both outcomes are service-use proxies rather than adjudicated syndromal relapse.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Cearns[55]	Core AI/ML: Direct relapse-relevant	BiDirect cohort with an acute MDE at baseline; complete multimodal data n=380 (102 rehospitalised, 278 not).	Self-reported rehospitalisation for an acute depressive episode during 2-year follow-up.	Repeated nested internal CV (inner feature/model selection; outer performance estimation). No external validation.	Multimodal AUC 0.677, balanced accuracy 0.631, sensitivity 0.575 and specificity 0.687; not significantly better than clinical-only prediction (p=0.15).	Does broad baseline fusion add useful information? Endpoint is a severe service-use proxy and biological gains were limited.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Habets[9]	Supporting: Adjacent modality-specific	NESDA participants with current MDD and/or dysthymia, n=804 overall; proteomic and clinical-plus-proteomic analyses n=611 (training/CV n=490; held-out test n=121).	Remitted versus non-remitted MDD/dysthymia at 2 years, not a discrete relapse event.	Repeated CV confined to training data, followed by a held-out test set. No external validation.	Test AUC 0.63 clinical-only versus 0.78 clinical plus proteomics (p=0.013); proteomics-only AUC 0.67 and balanced accuracy 0.68. Other omics layers did not add significant value.	Shows non-redundant multimodal gain for course status, but does not establish post-remission relapse prediction.

Vairavan[5]	Core AI/ML: Direct relapse-relevant	Two independent recurrent-MDD cohorts meeting a study-defined recovery criterion (MADRS ≤14): OBSERVEMDD n=227 (66 relapsed) and CBN-WELL n=50 (14 relapsed). Total 277 patients, 80 relapsed. Continuous actigraphy and weekly self-reports were collected throughout.	Relapse: confirmed MADRS/CGI-S worsening or specified clinical events (Table 1 of source). Followed for ≥1 year or until first relapse; clinic visits every 2 months. Evaluable actigraphy-linked visits: 50 relapse + 1,204 non-relapse in OBSERVEMDD; 10 + 146 in CBN-WELL. These visit counts differ from patient event counts.	Same N-of-1 framework evaluated in two independent cohorts with different devices; each patient supplied ≥2 months of non-relapse training data, followed by sequential testing and updating. Cross-cohort replication of the framework, not external testing of one fixed fitted model. No	OBSERVEMDD / CBN-WELL: balanced accuracy 73.5% / 71.0%; PPV 12.5% / 14.6%; NPV 98.3% / 97.2%; false alarms 1.8 / 2.3 per patient-year. Median lead 21 days (IQR 7–28) / 14 days (IQR 14–26.25) before the determining clinic visit. Table 3 reports no CIs for these performance estimates. Questionnaire burden reductions of 61.9% / 53.0% are retrospective estimates of assessments used, not observed deployment effects.	When does within-person deviation begin? Lead times depend on visit-based labels and use of current and subsequent weekly surveys; clinical confirmation remains necessary. Source locations: pp. 2–5, 8–10; Figure 1 and Tables 1–3.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
				prospective alert-triggered care pathway.		
Matcham[56]	Supporting: Contextual/methodological	RADAR-MDD feasibility cohort, n=623 adults with recurrent MDD across three European sites; smartphone and wearable streams.	Recruitment, retention and data availability over up to 24 months; no relapse-prediction endpoint.	Feasibility/descriptive analyses only; no prediction-validation design.	Only 110/623 participants achieved more than 50% data availability across all streams.	Can sustained between-visit monitoring be maintained? Informs feasibility, not predictive validity.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Matcham[47]	Supporting: Direct relapse-relevant	RADAR-MDD recurrent-MDD cohort; relapse analysis n=218 (624 assessments, 39 relapse assessments); median follow-up 541 days (IQR 401–730).	Subsequent depressive relapse; recurrent-MDD cohort not uniformly remitted at baseline. Every 3 months: CIDI-SF MDD criteria plus IDS-SR greater than 25 and remission within the previous 6 months, with telephone confirmation. Fitbit sleep features summarised over the preceding 4 weeks.	Longitudinal Bayesian association models with participant random effects; sensitivity analyses and 10-fold CV for interaction-model comparison. No external validation; no individual relapse classifier.	Sleep duration variability, fragmentation, lower efficiency, and midpoint variability track relapse risk. Per 1-SD within-person increase, sleep-time variability and midpoint variability each increased relapse probability by 2.8 percentage points (89% credible intervals 1.1–4.6 and 1.2–4.5). These are association estimates, not AUC or calibration.	Which passive sleep domains precede relapse? Direct prospective association, but not a validated alert model.

Tonon[46]	Supporting: Direct relapse-relevant	Canadian multisite treatment-responsive MDD cohort (MADRS ≤14). n=93 with valid actigraphy (Abstract, p. 379; Results, p. 382).	The descriptive Table reports 28 relapsed participants, although its subgroup counts total 94; no event proportion was derived. Panel-adjudicated relapse over 1–2 years: sustained MADRS ≥22 with CGI-S increase ≥2, hospitalisation, suicidal intent/behaviour or clinician-directed medication change. Visits every 8 weeks; only first relapse	Cox association models with baseline and time-varying predictors, within-participant clustering and sensitivity analyses. No external validation or train–test predictive evaluation. Figure 1 distinguishes baseline-MADRS adjustment from the model additionally adjusting for	Baseline HRs (Abstract, p. 379; Results, p. 382): sleep regularity 0.46 (95% CI 0.28–0.74); relative amplitude (RA) 0.45 (0.29–0.70). Concurrent-MADRS-adjusted RA HR 0.60 (95% CI 0.36–0.98; Abstract, p. 379; Results, pp. 382–383). Figure 1 (p. 383) reports the same interval as 0.37–0.97; this discrepancy within the source is left unresolved. Concordance 0.76 / 0.78 is apparent model discrimination, not calibration (Results, p. 383).	Do sleep/activity patterns precede relapse? Prospective associations do not establish a validated alert. Source locations: Abstract, p. 379; Methods, pp. 380–382; Results, pp. 382–383; Table and Figure 1, p. 383.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
			considered; preceding 2 weeks excluded.	time-varying MADRS.		

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Lee[45]	Supporting: Direct relapse-relevant	Longitudinal mood-disorder wearable cohort; MDD analytic subgroup n=95 (mixed-diagnosis parent cohort n=270).	Clinically identified major depressive episode within the next 3 days, using the preceding 18 days of passive circadian features.	Within-cohort rolling temporal evaluation using 18 days for training and the subsequent 3 days for testing in the same participant group; MDD subgroup reported separately. No external validation.	MDD subgroup: sensitivity 0.915, specificity 0.943, accuracy 0.938 and AUC 0.973; event dependence and same-cohort evaluation limit transportability.	Can passive rhythms warn of imminent recurrence? High internal performance requires prospective external replication.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Cho[44]; Kiguchi[48]	Supporting: Adjacent modality-specific	Cho: pooled bipolar/unipolar sample (n=55). Kiguchi: 89 recurrent-MDD patients in remission (BDI-II ≤9) on maintenance pharmacotherapy; 52-week follow-up. Available K6 outcomes: 3,303/4,628 patient-weeks; clusters n=30 and n=59.	Cho: pooled mood-state/episode transitions. Kiguchi: next-week K6 class using the preceding 4 weeks of features; thresholds K6 ≥5 and ≥9 for binary discrimination. No adjudicated MDE-relapse event count. K6 patient-weeks are repeated observations, not independent patients.	Cho: internal individual-level evaluation. Kiguchi: leave-one-participant-out CV; pragmatic simulation adds 2, 4, 8 or 12 weeks from the held-out patient and tests that patient's final 6 weeks (132 / 274 test records by cluster). No external validation.	Kiguchi standard CV AUCs (95% CIs), clusters 1 / 2: K6 ≥5, 0.89 (0.86–0.92) / 0.90 (0.88–0.92); K6 ≥9, 0.89 (0.87–0.92) / 0.91 (0.89–0.93). Standard weighted κ 0.41 (0.34–0.47) / 0.10 (0.06–0.13); at 12 weeks in pragmatic CV, 0.68 (0.55–0.81) / 0.59 (0.48–0.70). Twelve weeks is a tested duration, not an established minimum.	How much personal history helps symptom monitoring? Adjacent K6-state evidence, not MDD-relapse validity. Kiguchi source locations: pp. 2–4, sections 2.2–2.5 and 3.1–3.4; Tables 1–2 and Figure 2.

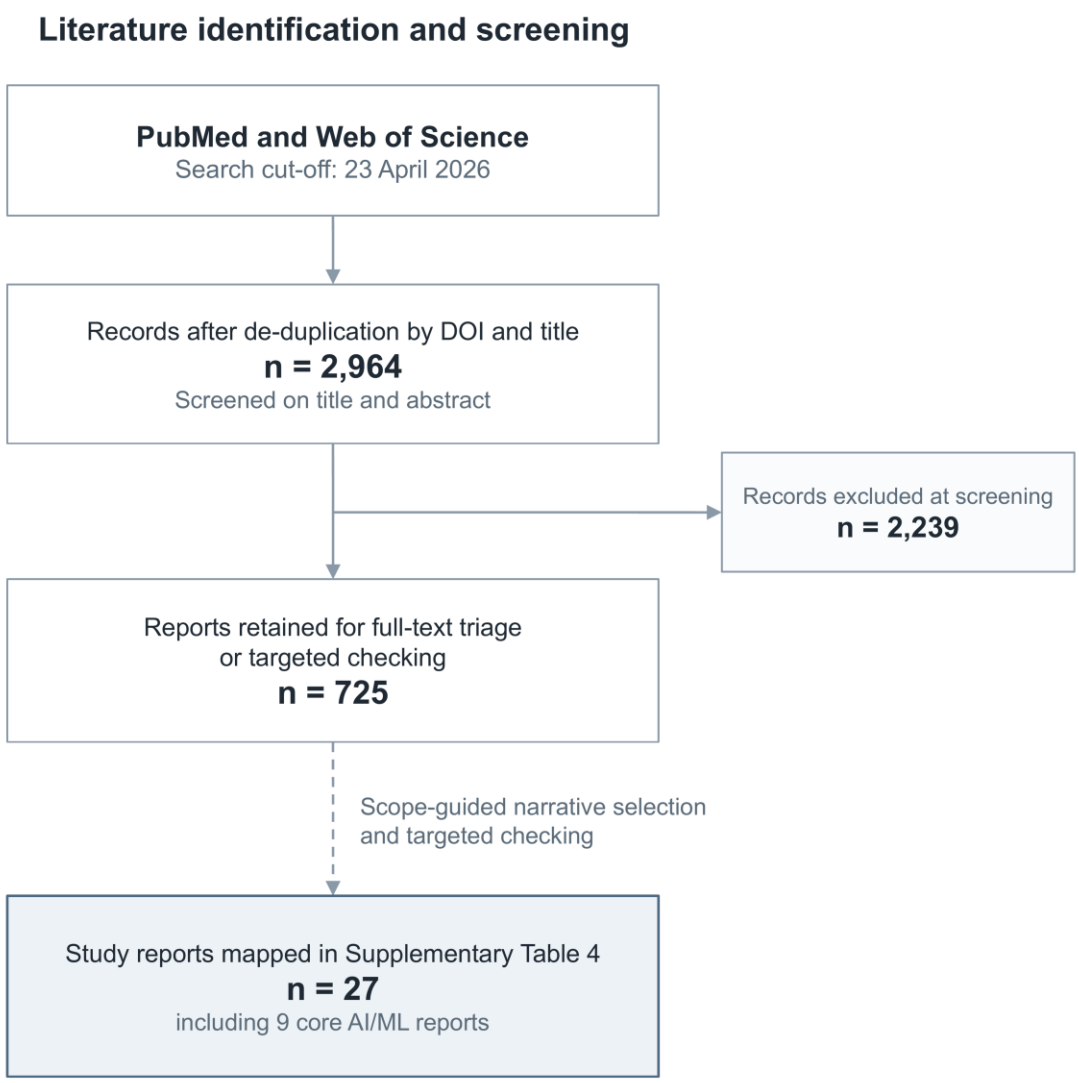
Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Di[10]	Supporting: Adjacent modality-specific	Non-standardised recordings from semi-structured diagnostic interviews in female recurrent-MDD cases and controls. Model training n=7,654; independent replication n=1,189 (490 cases, 699 controls). Many cases were in remission. Pitch-related acoustic features.	Cross-sectional recurrent-MDD case–control classification; associations with current depressive symptoms were examined separately in the replication sample. Neither analysis predicts relapse after remission.	Independent-sample replication.	Independent-replication AUC 0.90 for an XGBoost model combining pitch-related features with demographic and recording-quality covariates. The target was recurrent-MDD case–control status, not current depressive state or prospective relapse.	Can speech distinguish MDD cases from controls? Case–control discrimination is not prospective relapse prediction.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Menne[52]; Hartnagel[53]	Supporting: Adjacent modality-specific	Menne: n=96 (44 MDD, 52 controls); positive and negative narratives elicited with standardised, pre-recorded instructions. Hartnagel: n=22 inpatients, with 710 paired speech–symptom observations during sleep-deprivation therapy.	Menne: current MDD status and concurrent BDI-II scores; symptom correlations were analysed across patients and healthy controls combined. Hartnagel: concurrent symptom severity and longitudinal state generalisation; neither study evaluates post-remission relapse.	Internal subject-level and chronological validation. No external validation (no independent relapse cohort).	Menne: speech features were associated with concurrent BDI-II scores across the combined sample. Hartnagel: prediction of concurrent symptom severity failed to generalise to unseen patients under subject-level and chronological validation; more permissive splits performed better.	Where does speech-based state assessment generalise? Adjacent evidence for Confirm, not direct relapse prediction.

Studies	Core status / evidence directness	Population, sample and design	Outcome definition, events and follow-up	Validation	Performance and design convergence	Clinical question / evidence boundary
Othmani[50]; Eichstaedt[51]	Supporting: Adjacent modality-specific	Othmani: DAIC-WOZ audiovisual interviews (source dataset n=189). Eichstaedt: retrospective social-media language linked to EHR diagnosis codes.	Algorithmically constructed recurrence (Othmani) or first recorded depression diagnosis code (Eichstaedt), not observed post-remission relapse.	Retrospective/internal validation. No external validation.	Demonstrates state-sensitive audiovisual/language signal; reported metrics target non-equivalent proxy outcomes and are not compared with direct studies.	Method development only for this review question; endpoints do not establish prospective remitted-MDD relapse.

Note: Core status and evidence directness were coded separately. Core status identifies reports meeting the MDD-specific individual-level prediction criterion defined in Supplementary methods, whereas evidence directness follows the three descriptive endpoint groupings defined there. Supporting reports were retained where they provided direct prospective prognostic, modality-specific, feasibility, implementation, or contextual evidence relevant to the SMC framework. These labels are descriptive and are not risk-of-bias ratings or formal evidence grades. Endpoint heterogeneity prevents direct comparison of performance metrics across rows; each entry reports the original study's operational endpoint. Performance is reported only when the source evaluated an individual-level model; prospective or group-level associations are not converted into individual prediction evidence. Validation labels distinguish internal CV or held-out evaluation from independent-sample or cross-cohort validation; 'No external validation' means that no independent cohort, health system, or external setting was used for performance evaluation. Counts refer to the relevant analytic sample where available. AI: artificial intelligence; AUC: area under the receiver operating characteristic curve; BDI-II: Beck Depression Inventory-II; CGI-S: Clinical Global Impression-Severity; CI: confidence interval; CIDI-SF: Composite

International Diagnostic Interview-Short Form; CV: cross-validation; DAIC-WOZ: Distress Analysis Interview Corpus Wizard-of-Oz; DSM: Diagnostic and Statistical Manual of Mental Disorders; EHR: electronic health record; F1: F1 score; fMRI: functional magnetic resonance imaging; HAM-D: Hamilton Depression Rating Scale; HR: hazard ratio; IDS-SR: Inventory of Depressive Symptomatology-Self-Report; IQR: interquartile range; K6: six-item Kessler Psychological Distress Scale; LDA: linear discriminant analysis; LIFE: Longitudinal Interval Follow-up Evaluation; LOOCV: leave-one-out cross-validation; MADRS: Montgomery-Åsberg Depression Rating Scale; MDE: major depressive episode; MDD: major depressive disorder; NPV: negative predictive value; PPV: positive predictive value; RA: relative amplitude; rs-fMRI: resting-state functional magnetic resonance imaging; SCID: Structured Clinical Interview for DSM; SD: standard deviation; SMC: Stratify-Monitor-Confirm; SMOTE: Synthetic Minority Over-sampling Technique. Reference numbers correspond to the main-text reference list. For Puder[17], Vairavan[5], Kiguchi[48] and Tonon[46], page and table locators refer to the original reports. Reported CIs, fold-level SDs and IQRs are distinguished; missing CIs are not reconstructed. This map tabulates the 27 reports on which the modality-to-decision argument rests; other prediction studies cited in the main text are discussed there with their published estimates rather than mapped here, including traditional (non-AI/ML) clinical tools such as Judd[8] and Wang[28], and course-prediction models developed in currently depressed cohorts such as Schmaal[14] and Dinga[54].



Supplementary Figure 1 Literature identification and screening. This figure documents identification and screening for a critical narrative Review and is not PRISMA systematic-review reporting.