Supplementary File 4. Detailed Source-Level Extraction.

This appendix provides extracted data for all 40 included evidence sources, organized into seven tables: source characteristics, AI characteristics, mapping to the Knowledge-to-Action (KTA) cycle, implementation science theories, models, and frameworks (TMFs) and data science frameworks used, evaluation approaches and comparators, outcome domains reported, and risks and adverse effects.

Table A1. Source characteristics (N = 40).

Source	Country	Source type	Study design	Health / service domain	Setting
Bartholomew (2022)	USA	Primary research article	Quantitative; Observational (cross-sectional); Not applicable / not stated	Public health / policy implementation	Florida, USA; state-level surveillance data aggregated at ZIP Code Tabulation Areas (ZCTAs)
Barwise (2026)	USA	Primary research article	Quantitative; pragmatic two-arm stepped-wedge cluster randomized trial across 35 inpatient units	Inpatient care / interpreter services / health equity	Hospital inpatient units (EHR-integrated workflow with secure chat to bedside nurses)
Berkel (2019)	USA	Conceptual, methodological or review article	Not applicable; methodological synthesis with illustrative empirical measurement findings and a proposed future monitoring system	Prevention / community mental health (parenting program for divorcing families)	Community mental health agencies contracted through family courts in four Arizona counties, USA
Berkel (2023)	USA	Primary research article	Quantitative; Other; Not applicable / not stated	Healthcare (primary care) / implementation monitoring	Primary care settings delivering FCU4Health (type 2 hybrid effectiveness-implementation trial)
Brown (2013)	USA	Conceptual, methodological or review article	Not stated; Not applicable / not stated; Not applicable / not stated	Public health / HIV prevention	Community-based HIV prevention and behavioral intervention implementation
Chan (2025)	Australia	Study protocol	Mixed-methods; Other; Interviews; Focus groups; Ethnography/observation; Human-centred co-design + ethnography; usability testing; NLP/LLM development and performance evaluation; planned implementation evaluation vs standard approaches.	Health systems, implementation infrastructure (initial focus: cancer genetics/genomics)	Healthcare systems (initial focus: cancer genetics/genomics in Australia); intended for broader health settings
Chen (2025)	USA	Primary research article	Quantitative; quasi-experimental pre-post analysis with propensity-score matching and sensitivity analyses	Healthcare / health services	Clinics using EHR clinical decision support for tobacco cessation
Chin-Purcell (2025)	USA	Conference abstract	Mixed-methods; Other; Interviews; Qualitative interviews + quantitative agreement assessment (manual vs GPT-4 assisted coding).	Implementation science methods; public health needs assessment	County-level public health context (USA)
Creed (2022)	USA	Study protocol	Mixed-methods protocol: Phase I user-centred design, usability, and implementation-readiness evaluation; Phase II hybrid type 2 stepped-wedge cluster randomized trial	Mental health services / implementation	Community mental health agencies; CBT delivery and supervision/QA
Dabravolskaj (2025)	Canada	Primary research article	Quantitative; Other; Not applicable / not stated	Public health / school-based health promotion	Elementary schools in disadvantaged communities across Alberta, British Columbia, Manitoba, and Northwest Territories (APPLE Schools network), Canada
El-Bassel (2025)	USA	Framework development article	Not applicable; conceptual model development with a retrospective HCS use case and proposed post-hoc AI analyses	Community-engaged implementation science; public health (opioid overdose prevention)	Community settings (multi-community US implementation study use case)
Goedken (2025)	USA	Conference abstract	Mixed-methods; Observational (cross-sectional); Interviews; Structured interviews (n=12) + post-pilot	Implementation evaluation; health system AI initiatives (VA)	Veterans Health Administration (National AI Institute initiatives)

			surveys with open-ended questions (n=20); AI-assisted rapid qualitative analysis with human validation.		
Golz (2022)	Switzerland	Primary research article	Qualitative; Not applicable / not stated; Interviews	Mental health / psychiatry	Two psychiatric hospitals (German-speaking Switzerland)
Griffen (2024)	USA	Primary research article	Quantitative; Other; Not applicable / not stated	Autism / intellectual & developmental disabilities; hygiene skill acquisition	Privately owned ABA clinic (bathroom handwashing sessions)
Huguet (2024)	USA	Conceptual, methodological or review article	Not stated; Not applicable / not stated; Not applicable / not stated	Implementation science methods (cross-domain)	General (multiple healthcare/public health settings)
Ironside (2025)	USA	Conference abstract	Mixed-methods secondary analysis of two separate implementation trials (expert-human feedback in Trial 1; AI-based feedback in Trial 2), using multilevel logistic regression and post-period focus groups; not a direct randomized AI-versus-human comparison	Behavioral healthcare (motivational interviewing quality for substance use/mental health services)	Community behavioral healthcare organizations, USA; two trials comparing human expert vs AI-based MI feedback over 9 months
Jones (2022)	USA	Report	Comparative case-study report; one of five cases describes a planned AI-enabled social-media mining application	HIV (Ending the HIV Epidemic); implementation research methods	United States; HIV prevention/care settings (varied across cases)
King (2023)	USA	Conceptual, methodological or review article	Not stated; Not applicable / not stated; Not applicable / not stated	Critical care / intensive care (ICU)	Intensive care units (hospital critical care)
Knippen (2025)	USA	Primary research article	Mixed-methods; Before-after; Interviews; Hybrid implementation study: QI phase using iterative PDSA cycles (Model for Improvement); registry audit of process outcomes (quantitative); closing interviews analysed with constructivist reflexive thematic analysis supported by AI-assisted qualitative software; themes mapped to Normalization Process Theory (NPT) and CFIR.	Nutrition / medical nutrition therapy for gestational diabetes mellitus (GDM)	Five clinical sites in the USA (hospital outpatient clinics, maternal fetal medicine outpatient clinics, diabetes self-management clinic) providing MNT for GDM
Lee (2026)	USA	Primary research article	Quantitative; Other; Not applicable / not stated	Palliative care / serious illness communication (goals-of-care discussions)	University of Washington / UW Medicine hospitals, Seattle, WA, USA; two pragmatic clinical trials (PICSI Hospital Trials 1 & 2)
Li (2021)	USA	Primary research article	Mixed-methods methodological study: literature review; patient and family-caregiver focus groups; quantitative factor, latent-class, finite-mixture, and machine-learning analyses; and expert review	Transitional care / hospital discharge / value-based care	Hospitals implementing transitional care strategies
Lindamer (2022)	USA	Primary research article	Mixed-methods; Observational (cross-sectional); Other; Provider and patient surveys with closed-ended and open-ended questions; descriptive statistics plus NLP (n-gram analysis) of free-text responses.	Cognitive rehabilitation / traumatic brain injury / neuropsychiatric disorders	Real-world clinical settings (notably VA- or university-related clinics; some private practice)
Mac Aonghusa (2020)	Ireland	Prototype development article	Methodological development using expert-annotated documents, iterative algorithm training/testing, and human checking of predictions	Public health / behavioral science	Human Behaviour-Change Project (HBCP) knowledge system for behavior change intervention evidence
Michie (2017)	United Kingdom	Study protocol	Not stated; Not applicable / not stated; Not applicable / not stated	Behaviour change interventions; evidence synthesis (initial focus: smoking cessation)	Evidence synthesis and decision support (knowledge system / user interface)
Moussa (2021)	Australia	Primary research article	Mixed-methods; Other; Interviews; Implementation barriers were recorded by facilitators during a 2-year facilitation program; strategies were categorized; a	Community pharmacy services / implementation of professional pharmacy services	Community pharmacies (Health Destination Pharmacy program)

			random forest model was trained to predictstrategy success		
Pearson (2020)	USA	Conceptual, methodological or review article	Not stated; Not applicable / not stated; Not applicable / not stated	Cardiovascular/HLBS disorders; precision medicine & precision public health	Clinical medicine and public health; learning health systems/communities (workshop-based agenda)
Rappon (2021)	Canada	Primary research article	Quantitative; Other; Not applicable / not stated	Sustainability of evidence-based interventions (cross-sector)	Web-based decision support tool (sustainability planning)
Rinne (2024)	USA	Primary research article	Mixed-methods; Observational (cross-sectional); Other; Mixed-methods methodological study: expert panel discussion (qualitative) summarised via ChatGPT-3.5; Likert + open-ended participant survey evaluating the AI-generated manuscript; content analysis of ChatGPT-generated manuscript vs. human-revised version.	Implementation science methods / health services research (VA)	Veterans Affairs (VA) Quality Enhancement Research Initiative (QUERI) expert panel session, USA
Rosen (2025)	USA	Primary research article	Quantitative; RCT; Not applicable / not stated	Mental health / PTSD psychotherapy (military treatment facilities)	Eight U.S. military treatment facilities (four Army, three Air Force, one Navy), USA
Rosen (2026)	Germany	Primary research article	Qualitative; Not applicable / not stated; Focus groups	Physical therapy / clinical practice guideline adherence (low back pain and COPD)	German university hospital inpatient physiotherapy (TUM University Hospital Rechts der Isar, Munich) and outpatient physiotherapy practices, Germany
Shafran (2026)	UK	Conceptual, methodological or review article	Not stated; Not applicable / not stated; Not applicable / not stated	Mental health / cognitive behavioural therapy training and supervision	Cross-cutting / international (UK, USA, Sweden, Korea, Australia, Canada, India authors); focus on CBT training/supervision globally including low- and middle-income countries
Shankar (2025)	Singapore	Primary research article	Mixed-methods; Before-after; Other; Integrated Knowledge Discovery in Databases (KDD) process + action research cycles: retrospective analysis of 126,134 patient feedback entries using a suite of NLP techniques, with three stakeholder workshops (n=28) to co-develop and monitor five targeted interventions (pre-post comparison of self-reported outcomes).	Health system quality improvement (patient experience/feedback analytics)	Alexandra Hospital, National University Health System, Singapore; inpatient, outpatient, and emergency care settings (2023)
Sibley (2025)	USA	Primary research article	Mixed-methods randomized controlled trial: therapists were randomized within agency and adolescent clients were independently randomized to AI-assisted or standard supports; quantitative service-delivery and fidelity outcomes were supplemented by supervisor interviews and an end-of-study focus group	Child/adolescent mental health (ADHD CBT with motivational interviewing)	Three community mental health agencies serving ethnic/racial minority and low-income youth in a large culturally diverse southeastern U.S. city, USA
Stanfield (2024)	USA	Primary research article	Quantitative; Simulation; Not applicable / not stated	Cancer screening / preventive care (colorectal cancer)	Population-based colorectal cancer screening programs / public health planning
Swanson (2025)	USA	Primary research article	Mixed-methods; Observational (cross-sectional); Document analysis; Content analysis of 108 consultation notes + program metrics; ChatGPT Plus used to facilitate thematic synthesis with iterative validation.	Implementation science capacity building / consultation services	University-based DIS capacity building programs (UC San Diego)
Trinkley (2024)	USA	Conceptual, methodological or review article	Not stated; Not applicable / not stated; Not applicable / not stated	Implementation science methods / learning health systems	Cross-cutting (health systems and research contexts)
Wang (2016)	USA	Primary research article	Quantitative; Other; Not applicable / not stated	Mental health / child welfare implementation (foster care system)	County-level implementation of evidence-based mental health intervention in public service systems

Ye (2025)	USA	Primary research article	Quantitative observational cross-sectional analysis with practice-facilitator stakeholder review of selected features and interpretation	Primary care quality improvement / practice facilitation	Primary care/outpatient clinics (practice facilitation program; health system QI context)
Younas (2024)	Canada	Conceptual, methodological or review article	Not stated; Not applicable / not stated; Not applicable / not stated	Implementation science (conceptual commentary) / nursing	Cross-cutting (healthcare practice and policymaking contexts)
Zhong (2025)	China	Framework development article	Mixed-methods framework development: systematic review and meta-aggregation; stakeholder consultation; ChatGPT-assisted classification; external validation of the CFGI	Clinical practice guideline development and implementability (general and Traditional Chinese Medicine)	Mixed-methods framework development spanning mainland China (east, central, west regions) stakeholder consultations and the 2024 Global Implementation Society & Nigeria Implementation Science Alliance Conference in Abuja, Nigeria

Table A2. AI approach, methods, models, training, and maturity (N = 40).

Source	AI category	AI methods (free text)	Models / tools	Training design	Maturity
Bartholomew (2022)	Conventional (non-deep) machine learning	Poisson LASSO regression applied to surveillance and hospitalization indicators to identify predictors of acute HCV infection and map high-priority locations for syringe services program implementation; Random Forest was used as a sensitivity/specification analysis.	Poisson LASSO regression; Random Forest	Supervised regression and variable selection with 10-fold cross-validation	Prototype (research)
Barwise (2026)	Conventional (non-deep) machine learning	A pre-existing EHR-integrated predictive model generated a daily complexity-ranked list of adult inpatients with non-English language preference. Language services reviewed the list and sent secure-chat nudges to bedside nurses to promote in-person interpreter use. A companion interim abstract (Strechen 2024) reported the same trial at its midpoint and is treated as supporting information, not as a separate source.	Control Tower workflow; predictive palliative-care/medical-complexity score; EHR-derived indicators (length of stay, level of care, procedures, events, and clinical notes); secure-chat nudge	Pre-existing supervised predictive model; training and model-validation details were not reported in the trial paper or companion interim abstract. The randomized trial evaluated the AI-enabled workflow rather than model performance.	Deployed in trial; continued after trial completion
Berkel (2019)	Not applicable (conceptual, methodological or review article discussing multiple computational approaches)	Methodological synthesis and proposed future implementation-monitoring system. The paper reports illustrative validation findings from related measures and prior computational studies, but does not evaluate one new integrated AI system.	Proposed custom text and speech classifiers; web-based data collection and feedback system	Not applicable	Not applicable
Berkel (2023)	Multiple AI approaches compared: conventional (non-deep) machine learning; non-generative deep learning	Two distinct NLP/ML pipelines were tested to predict fidelity and competent adherence from English- and Spanish-language session transcripts: TF-IDF features with support-vector regression and multilingual BERT embeddings with an LSTM model.	TF-IDF + support-vector regression; multilingual BERT + LSTM; COACH fidelity measure	Supervised prediction of observer-rated COACH scores; validation against observer ratings and engagement outcomes	Prototype (research)
Brown (2013)	Not applicable (conceptual, methodological or review article discussing multiple computational approaches)	Conceptual article illustrating potential uses of mobile technologies, supervised ML, computational linguistics, speech recognition, social-network analysis, agent-based models, intelligent data analysis, and systems engineering. These were not integrated or evaluated as one AI application in this paper.	Mobilyze (named example); computational linguistics; social-network analysis; agent-based models	Not applicable	Not applicable
Chan (2025)	Multicomponent AI system: generative AI (LLM-based); non-generative deep learning; conventional (non-deep) machine learning; ontology-based knowledge representation	ImpleMATE combines an implementation ontology, locally hosted LLMs, BERT-based named-entity recognition and relation extraction, topic modelling, ontology-based retrieval, and LLM synthesis to extract implementation concepts and provide context-specific decision support.	ImpleMATE; implementation ontology; BERT; GPT/Mistral/LLaMA; LDA; BERTopic; ontology-based retrieval	Planned supervised fine-tuning and prompt-based use; expert-annotated ground truth; reserved internal validation subset; performance comparisons across models and prompts	Prototype / under development
Chen (2025)	Conventional (non-deep) machine learning	A structural topic model inferred EHR-use activities from audit-log sessions before and after implementation of a tobacco-cessation clinical decision-support alert. Four domain experts labelled 14 topics; automatic validation examined expected CDS-related patterns, and logistic-regression and Random-Forest models tested whether topic distributions distinguished before- from after-check-in sessions. A 2023 companion abstract reported an earlier 12-topic developmental analysis and is treated as supporting information, not as a separate source.	Structural topic model (R stm); LDA developmental comparator; logistic regression; Random Forest; Kneedle topic-number selection; propensity-score matching and IPWRA	Unsupervised STM trained on 3445 encounters; topic number selected using semantic coherence and exclusivity; iterative labelling by four domain experts; automatic construct validation; held-out classification testing; propensity-score matching and sensitivity analyses	Prototype (research)
Chin-Purcell (2025)	Generative AI (LLM-based)	GPT-4 was prompted to apply a qualitative codebook, identify relevant text segments, provide coding rationales, and generate downstream qualitative analysis for comparison with manual coding.	GPT-4; Atlas.ti for manual coding	Prompt-based use of a pretrained LLM; no fine-tuning reported	Prototype (research)

Creed (2022)	Non-generative deep learning	LyssnCBT combines speech-signal processing, voice activity detection, speaker diarization, automated speech recognition, and transformer-based deep neural networks to estimate 11 Cognitive Therapy Rating Scale codes and generate therapist feedback through a cloud-based interface.	Lyssn / LyssnCBT; automated speech recognition; speaker diarization; transformer-based deep neural networks; Cognitive Therapy Rating Scale metrics; cloud-based feedback reports	Prototype models developed from 2494 expert-rated sessions using cross-validation and a distinct 30% test set; performance benchmarked against human reliability, with continuous in-house calibration. The protocol adds user-centred refinement, usability/readiness testing, and a planned stepped-wedge trial.	Real-world deployment and evaluation
Dabravolskaj (2025)	Conventional (non-deep) machine learning	Two supervised k-nearest-neighbours classifiers used tokenisation, stemming, and TF-IDF vectors to classify action-plan items as intervention activities and implementation strategies.	k-nearest neighbours (k=5); TF-IDF; RapidMiner 10.1	Supervised classification trained on 600 manually labelled items and applied to the remaining items; precision and recall exceeded 80% for true positives	Prototype (research)
El-Bassel (2025)	Not applicable (framework development)	The PRISM-Capabilities paper maps a portfolio of possible AI analyses to a new conceptual model. AI was not used during the HCS implementation; post-hoc analyses are proposed, and no single integrated AI system is specified.	BERT; BERTopic; GPT/RAG; Random Forest; XGBoost; Bayesian/causal models; GNNs; LSTM; sentiment and simulation approaches	Varies by proposed analysis; no single training design	Not applicable
Goedken (2025)	Generative AI (LLM-based)	VA GPT(Beta), a firewall-hosted generative AI tool, was prompted to organize, categorize, identify themes, generate insights, and summarize structured-interview and open-ended survey data from two VA National AI Institute pilots. Human reviewers checked all AI outputs against the raw datasets.	VA GPT(Beta); Microsoft Teams transcripts; Microsoft Forms survey data	Prompt-based use of a pretrained VA-hosted LLM; prompts were selected and refined iteratively; no fine-tuning or model training was reported	Pilot in real setting
Golz (2022)	Non-learning natural language processing and text analytics	Classical text processing included lemmatisation, word-frequency and n-gram analyses, and lexicon-based sentiment analysis using the German SentiWS resource.	R/RStudio; spacyr; German spaCy model; SentiWS lexicon	No ML training; lexicon- and frequency-based analysis	Prototype (research)
Griffen (2024)	Symbolic or rules-based AI	GAINS provided real-time, stepwise audio/visual guidance using pre-programmed adaptive logic based on therapist-entered learner responses and recorded performance data.	GAINS (Guidance, Assessment, and Information Systems)	No learning or model training reported; pre-programmed adaptive rules	Prototype in real setting
Huguet (2024)	Not applicable (conceptual, methodological or review article discussing multiple computational approaches)	Viewpoint describing potential applications of supervised and unsupervised ML, clustering, deep learning, reinforcement learning, ensemble learning, NLP, predictive decision support, and model recalibration across implementation stages. No single AI application was conducted.	FINDER (named example); multiple ML and NLP approaches	Not applicable	Not applicable
Ironside (2025)	AI family not reported	The abstract analyzed determinants of provider engagement with automated AI-based motivational-interviewing feedback in one implementation trial and contrasted patterns descriptively with a separate trial offering expert-human feedback. The AI platform and model architecture were not reported.	AI-based motivational-interviewing feedback tool (not named); separate expert-human feedback tool in companion trial	Pretrained system used in practice; model architecture and training design not reported	Pilot in real setting
Jones (2022)	AI family not reported	Case Study 3 describes planned text-based analysis and development of a social-media classification/targeting algorithm to identify a virtual cohort, assess PrEP barriers, and inform strategy tailoring. The specific model architecture was not reported.	Social-media text classification/targeting algorithm (not specified)	Not stated	Under development / not stated
King (2023)	Not applicable (conceptual, methodological or review article discussing multiple computational approaches)	Narrative review describing reinforcement-learning systems, recommender systems, multi-armed bandits, ML-triggered prompts, dynamic checklists, EHR mining, NLP, deep learning, and eye-tracking plus ML. No single application was developed or tested.	AI Clinician (named example); multiple data-science approaches	Not applicable	Not applicable

Knippen (2025)	Generative AI (LLM-based)	AILYZE generated an inductive thematic analysis and a second analysis using the researcher-developed codebook; the research team compared both AI analyses with three rounds of manual reflexive thematic analysis. ChatGPT-4 was used only to refine theme labels and descriptions, with final wording determined by the investigator.	AILYZE; ChatGPT-4; manual coding in Microsoft Word and Excel	Prompt- and codebook-based use of commercial pretrained generative-AI tools; no fine-tuning or model training; AI analyses were used for triangulation and reflexivity rather than as the primary analysis	Commercial AI products
Lee (2026)	Non-generative deep learning	A fine-tuned BioClinicalBERT transformer classifier screened EHR passages for documented goals-of-care discussions; screen-positive passages were routed to a REDCap-based human-adjudication workflow.	Fine-tuned BioClinicalBERT; REDCap; dynamic pooling	Supervised fine-tuning on 4,642 manually labelled clinical notes; fine-tuning is treated as the training approach, not a separate technology category	Production / deployed in two pragmatic trials
Li (2021)	Conventional (non-deep) machine learning	Random Forest explored the relative predictive importance of individual transitional-care strategies, strategy groups, and patient/system covariates for 30-day readmission; boosted trees were used as a sensitivity analysis. Factor analysis, latent-class analysis, finite-mixture modelling, and expert review were the primary strategy-grouping methods.	Random Forest; boosted trees; exploratory factor analysis; latent-class analysis; finite-mixture modelling	Supervised prediction using claims outcomes; boosted-tree sensitivity analysis; no train/test performance metrics reported. Machine learning was supplementary to statistical grouping and expert review.	Prototype (research)
Lindamer (2022)	Non-learning natural language processing and text analytics	Python-based n-gram frequency analysis was used to summarise common free-text survey responses about real-world implementation of CogSMART/CCT.	Python n-gram analysis	No ML training; frequency-based text analysis	Prototype (research)
Mac Aonghusa (2020)	Multicomponent AI system: NLP-based information extraction and ML-based outcome prediction (model families not reported); ontology-based knowledge representation; automated reasoning	The Human Behaviour-Change Project Knowledge System combines an ontology, expert-annotated NLP/statistical extraction, statistical matching, ML outcome prediction, and reasoning algorithms to synthesise trial reports and predict intervention outcomes.	HBCP Knowledge System; behaviour-change ontology; NLP extraction; ML prediction; reasoning algorithms	Supervised learning from expert annotations with iterative development, testing, human checking, and statistical matching	Prototype / methodological development
Michie (2017)	Multicomponent AI system: NLP-based information extraction and ML-based outcome prediction (model families not reported); ontology-based knowledge representation; automated reasoning	Protocol for a Knowledge System integrating a Behaviour Change Intervention Ontology, automated literature searching and NLP feature extraction, supervised ML prediction, logic/reasoning algorithms, and user/machine interfaces.	HBCP Knowledge System; Behaviour Change Intervention Ontology; NLP extraction; ML prediction; automated reasoning	Planned supervised annotation/extraction and prediction plus ontology-based reasoning	Protocol / planned
Moussa (2021)	Conventional (non-deep) machine learning	Random Forest classification predicted which facilitation strategies would resolve specific implementation barriers and generated predictive resolution percentages for barrier-strategy combinations.	Random Forest	Supervised classification with 10-fold cross-validation	Prototype (research)
Pearson (2020)	Not applicable (conceptual, methodological or review article discussing multiple computational approaches)	State-of-the-art review spanning linear and rule-based models, decision trees, Bayesian networks, ensemble trees, neural networks, deep learning, model stacking, ontologies, and computable phenotypes. No single AI application was conducted.	Multiple predictive-analytics approaches and external examples	Not applicable	Not applicable
Rappon (2021)	Conventional (non-deep) machine learning	A supervised Kohonen/XY-fused self-organising map was used to identify patterns linking sustainability determinants to outcomes and support tailored recommendations. Although it is an artificial neural network, it is not a deep-learning architecture.	Kohonen / XY-fused supervised self-organising map; R kohonen package	Supervised shallow neural-network modelling with leave-one-out cross-validation	Prototype (research)
Rinne (2024)	Generative AI (LLM-based)	ChatGPT-3.5 generated a scientific-manuscript draft from expert-panel breakout notes; participants evaluated and extensively revised the LLM-generated text.	ChatGPT-3.5 (May 25, 2023 version)	Prompt-based use of a pretrained LLM; no fine-tuning	Commercial product used in research

Rosen (2025)	Symbolic or rules-based AI	Apache cTAKES annotated negation, uncertainty, and other contextual elements in psychotherapy notes; validated rule-based decision logic classified sessions as prolonged exposure, cognitive processing therapy, or EMDR for trial outcome ascertainment.	Apache cTAKES 4.0.0; validated rule-based decision logic	No model training; decision rules were validated against 200 human-coded notes and reported sensitivity and specificity for each psychotherapy type	Deployed in trial
Rosen (2026)	Generative AI (LLM-based)	GAEP used GPT-4 Turbo to reformulate queries and generate concise, conversational answers grounded in two German clinical practice guidelines, while displaying underlying recommendations and references.	GAEP; OpenAI GPT-4 Turbo	Retrieval-/prompt-based use over curated guideline content; no fine-tuning reported	Prototype (research)
Shafran (2026)	Not applicable (conceptual, methodological or review article discussing multiple computational approaches)	Conceptual and guidance paper reviewing LLMs, chatbots, AI-driven virtual patients, multimodal simulation, automated fidelity feedback, multiagent LLM supervision, reasoning-augmented models, and other named systems. No original AI application was tested.	Named examples include Lyssn tools, Therapy Trainer, Virtual Insomnia Patients, Socrates 2.0, and Therabot	Not applicable	Not applicable
Shankar (2025)	Multicomponent AI system: conventional (non-deep) machine learning; non-learning natural language processing and text analytics	An operational multicomponent text-mining pipeline combined TF-IDF, n-grams, POS tagging, TextBlob sentiment analysis, LDA topic modelling, lexicon-based emotion detection, hybrid rule/ML text classification, aspect-based sentiment analysis, and named-entity recognition. Stakeholder workshops validated findings and translated them into five quality-improvement initiatives.	TextBlob; NLTK; Gensim/LDA; spaCy; NRC Emotion Lexicon; LIWC; TF-IDF; hybrid rule-based/ML classifier; comparative fine-tuned healthcare BERT models	Mixed unsupervised, lexicon/rule-based, and supervised components; topic coherence/perplexity assessment; hybrid-classifier accuracy assessment; manual review of 1000 sentiment classifications; stakeholder validation; comparative transformer analysis	Production / deployed
Sibley (2025)	Non-generative deep learning	Lyssn analyzed therapy-session audio to generate automated motivational-interviewing integrity feedback displayed through the Care4 clinical dashboard. The RCT also displayed human-coded STAND content-fidelity feedback. Model architecture was not fully reported in this paper; based on the platform descriptions cited in the article and prior verification of Lyssn, the AI was coded as non-generative deep learning.	Lyssn automated MI fidelity platform; transformer-based/deep-learning behavioral classifiers; MITI/MISC-derived fidelity metrics; Care4 clinical dashboard	Pretrained supervised models were developed and validated against expert human coding in prior Lyssn studies; no new AI-model training or technical validation was conducted in this RCT	Commercial product used in research
Stanfield (2024)	Conventional (non-deep) machine learning	Random Forest surrogate metamodels were trained on colorectal-cancer simulation runs to generate rapid predictions across intervention and population scenarios; linear and polynomial regression models were comparators.	Random Forest metamodels (500 trees); linear and polynomial regression comparators	Supervised training on simulation scenarios with held-out performance evaluation	Prototype (research)
Swanson (2025)	Generative AI (LLM-based)	ChatGPT Plus facilitated thematic synthesis, category labelling, code-frequency summaries, and organization of manually coded consultation-note data. The team used iterative prompting, detected and corrected counting and categorization errors, and reviewed all outputs.	ChatGPT Plus; manually coded consultation-note database; structured prompts and iterative refinement prompts	Prompt-based use of a pretrained LLM; no fine-tuning. Manual coding preceded AI-assisted synthesis, and the study team repeatedly validated and refined outputs.	Commercial product used in research
Trinkley (2024)	Not applicable (conceptual, methodological or review article discussing multiple computational approaches)	Debate paper discussing chatbots, NLP and automated transcription, generative AI, predictive ML, computer vision, and causal/mechanistic analyses as potential implementation-science tools. No single application was evaluated.	Multiple illustrative approaches and external examples	Not applicable	Not applicable
Wang (2016)	Conventional (non-deep) machine learning	Semi-supervised non-negative matrix factorisation classified implementation-broker communication logs into implementation stages; SVM and Naive Bayes were comparison models, and probabilistic change detection estimated stage transitions.	Semi-supervised NMF; linear SVM; Naive Bayes; probabilistic change detection	Semi-supervised learning using a human-labelled subset for training and validation	Prototype (research)
Ye (2025)	Conventional (non-deep) machine learning	The authors implemented classical statistical methods within a machine-learning framework: PCA for data-driven feature selection, parallel analysis and Velicer MAP testing	Principal component analysis; parallel analysis; Velicer MAP test; structural	Data-driven feature discovery and iterative refinement; internal-consistency and reliability testing;	Prototype (research)

		for feature-set optimization, stakeholder review for knowledge-driven feature validation, and structural equation modelling to estimate relationships among determinants, facilitation strategies, and outcomes.	equation modelling; RMSEA model-fit assessment	stakeholder construct validation; SEM model-fit assessment. No train/test prediction validation was performed.	
Younas (2024)	Not applicable (conceptual, methodological or review article discussing multiple computational approaches)	Commentary describing potential uses of predictive analytics, ML, NLP, expert systems, robotics, automated fact extraction, and adaptive checklists across the implementation continuum. No original AI system was evaluated.	Multiple potential and externally cited approaches	Not applicable	Not applicable
Zhong (2025)	Generative AI (LLM-based)	ChatGPT-4.0 classified candidate guideline-implementability constructs and proposed category titles; the research team subsequently refined the output using stakeholder feedback, literature triangulation, and internal framework logic.	ChatGPT-4.0	Prompt-based use of a pretrained LLM; no fine-tuning	Commercial product used as a research aid

Table A3. Mapping to the Knowledge-to-Action (KTA) cycle (N = 40).

Each cell indicates whether the source addressed the corresponding KTA step (Y = Yes; U = Unclear; blank = No). KTA1 = Synthesise implementation-relevant knowledge; KTA2 = Identify/prioritise problem or evidence-based intervention; KTA3 = Adapt knowledge to local context; KTA4 = Assess barriers and facilitators; KTA5 = Select and tailor implementation strategies; KTA6 = Implement and monitor implementation; KTA7 = Evaluate impact; KTA8 = Sustain and scale.

Source	KTA1	KTA2	KTA3	KTA4	KTA5	KTA6	KTA7	KTA8
Bartholomew (2022)		Y			Y			
Barwise (2026)					Y	Y	Y	
Berkel (2019)					Y	Y	Y	Y
Berkel (2023)						Y	Y	Y
Brown (2013)	Y	Y	Y	Y	Y	Y	Y	Y
Chan (2025)	Y	U	Y	Y	Y	Y	Y	Y
Chen (2025)						Y	Y	
Chin-Purcell (2025)		Y		Y				
Creed (2022)			Y	Y	Y	Y	Y	Y
Dabravolskaj (2025)			Y		Y			Y
El-Bassel (2025)	U	U	U	U	U	U	U	U
Goedken (2025)				Y			Y	
Golz (2022)				Y				
Griffen (2024)			U		U	Y	Y	
Huguet (2024)		Y	Y	Y	Y	Y	Y	Y
Ironside (2025)				Y		Y	Y	
Jones (2022)		Y		Y	Y	U	U	U
King (2023)	Y		Y	U	Y	Y	U	U
Knippen (2025)				Y			Y	
Lee (2026)						Y	Y	
Li (2021)					Y		Y	
Lindamer (2022)			U	U	U	Y	U	U
Mac Aonghusa (2020)	Y				Y		Y	
Michie (2017)	Y				Y			
Moussa (2021)				Y	Y	U	U	
Pearson (2020)				U	U	U	Y	Y
Rappon (2021)				Y	Y			Y
Rinne (2024)	Y			Y	Y	Y	Y	Y
Rosen (2025)							Y	
Rosen (2026)	Y			Y	Y			

Shafran (2026)	Y		Y	Y	Y	Y		
Shankar (2025)		Y		Y	Y	Y	Y	U
Sibley (2025)				Y	Y	Y	Y	
Stanfield (2024)		Y					Y	
Swanson (2025)							Y	
Trinkley (2024)	Y	Y	Y	Y	Y	Y	Y	Y
Wang (2016)				Y	U	Y		Y
Ye (2025)				Y	Y		Y	
Younas (2024)	Y	Y	Y	Y	Y	Y	Y	Y
Zhong (2025)	Y							

Table A4. Implementation science theories, models and frameworks (TMFs), adjacent frameworks, and AI/data science frameworks (N = 40).
Frameworks are separated into implementation science TMFs; adjacent behavioural, clinical, or quality-improvement frameworks; and AI/data-science frameworks, processes, or ontologies. "None / not stated" indicates that no framework in that category was explicitly invoked.

Source	Implementation science TMF(s)	Adjacent behavioural, clinical, or quality-improvement framework(s)	AI or data science framework, process, or ontology
Bartholomew (2022)	None / not stated	None / not stated	None / not stated
Barwise (2026)	None / not stated	None / not stated	None / not stated
Berkel (2019)	Durlak & DuPre implementation dimensions; theoretical Implementation Cascade Model	None / not stated	None / not stated
Berkel (2023)	Implementation Cascade Model	None / not stated	None / not stated
Brown (2013)	None / not stated	None / not stated	None / not stated
Chan (2025)	CFIR; TDF; ERIC; Learning Health System; Proctor implementation outcomes	None / not stated	Implementation-focused ontology / Learning Implementation System ontology (planned)
Chen (2025)	None / not stated	None / not stated	None / not stated
Chin-Purcell (2025)	None / not stated	None / not stated	None / not stated
Creed (2022)	Proctor implementation outcomes (acceptability, appropriateness, feasibility)	User-centred design; deliberate practice model; Cognitive Therapy Rating Scale	None / not stated
Dabravolskaj (2025)	Core functions and forms framework	Behaviour Change Wheel; Health Promoting Schools implementation components	None / not stated
El-Bassel (2025)	PRISM; RE-AIM	Capabilities Approach	None / not stated
Goedken (2025)	None / not stated	None / not stated	None / not stated
Golz (2022)	None / not stated	Technology Acceptance Model	None / not stated
Griffen (2024)	None / not stated	None / not stated	None / not stated
Huguet (2024)	Strategic Implementation Framework; EPIS; RE-AIM	None / not stated	None / not stated
Ironside (2025)	None formally named	Implementation climate and provider/supervisor attitudes toward motivational interviewing	None / not stated
Jones (2022)	EPIS; ERIC; Proctor implementation outcomes; RE-AIM reach	None / not stated	None / not stated
King (2023)	None / not stated	None / not stated	None / not stated
Knippen (2025)	CFIR; Normalization Process Theory	Model for Improvement / Plan-Do-Study-Act	None / not stated
Lee (2026)	None / not stated	None / not stated	None / not stated
Li (2021)	None / not stated	None / not stated	None / not stated
Lindamer (2022)	RE-AIM	None / not stated	None / not stated
Mac Aonghusa (2020)	None / not stated	None / not stated	Behaviour Change Intervention Ontology; HBCP Knowledge System
Michie (2017)	None / not stated	None / not stated	Behaviour Change Intervention Ontology; HBCP Knowledge System
Moussa (2021)	CFIR; TDF; TICD; Generic Implementation Framework	None / not stated	None / not stated

Pearson (2020)	None / not stated	None / not stated	None / not stated
Rappon (2021)	Dynamic Sustainability Framework; ERIC	None / not stated	None / not stated
Rinne (2024)	Proctor implementation outcomes; ERIC	None / not stated	None / not stated
Rosen (2025)	i-PARIHS; CFIR; ERIC-derived strategies	TACTICS rubric	None / not stated
Rosen (2026)	None / not stated	None / not stated	None / not stated
Shafran (2026)	None / not stated	CBT adherence/competence frameworks; WHO-UNICEF EQUIP	Fairness Metrics Ontology
Shankar (2025)	None / not stated	Action research cycles	Knowledge Discovery in Databases process; LIWC lexicon
Sibley (2025)	None / not stated	Innovation Tournament; STAND manualised CBT-MI protocol; Motivational Interviewing framework	None / not stated
Stanfield (2024)	None / not stated	None / not stated	None / not stated
Swanson (2025)	Core Functions and Forms framework; DIS competencies	None / not stated	None / not stated
Trinkley (2024)	None / not stated	None / not stated	None / not stated
Wang (2016)	Stages of Implementation Completion	None / not stated	None / not stated
Ye (2025)	Implementation Research Logic Model; CFIR	Participatory Impact Pathways Analysis	None / not stated
Younas (2024)	None / not stated	None / not stated	None / not stated
Zhong (2025)	Comprehensive Framework for Guideline Implementability (developed in the study)	GLIA; GUIDE-IT; GUIDE-M; AGREE; GLAFI (reviewed)	None / not stated

Table A5. Evaluation approaches and comparators (N = 31 application-focused sources).

Restricted to the 31 sources that described, developed, used, tested, implemented, or evaluated a specific AI system or prototype. Evaluation fields include completed or prospectively specified assessments. AI-specific evaluation includes internal or external validation, cross-validation, error analysis, fairness/bias analysis, explainability, and other model-quality assessments. Human-centred evaluation includes usability, user experience, satisfaction, stakeholder validation, or human-factors assessment.

Source	Study approach	AI-specific evaluation	Human-centred evaluation	Comparator	Comparator type / reference
Bartholomew (2022)	Quantitative	Cross-validation; internal hold-out validation; alternative-model specification check	None / not stated	Yes	Another AI model (Random Forest)
Barwise (2026)	Quantitative	None / not stated (trial evaluated the AI-enabled workflow rather than model performance)	Prior stakeholder interviews and community advisory feedback informed the workflow; nurse-perception evaluation was planned but not reported in the trial	Yes	Standard care relying on clinician-initiated interpreter requests
Berkel (2023)	Quantitative	Cross-validation	None / not stated	Yes	Standard manual process; Human (researcher/clinician); Another AI model; Non-AI technology
Chan (2025)	Mixed-methods	Internal validation; error/hallucination analysis; semantic-equivalence assessment; model/prompt performance comparison; bias monitoring planned but not operationally specified	Usability testing; human factors/UX evaluation; stakeholder co-design	Yes	Expert-annotated ground truth / manual process; human expert review; alternative AI models and prompts
Chen (2025)	Quantitative	Expert topic review; automatic construct validation; held-out classification performance; sensitivity analyses; alternative-model comparison	None / not stated	Yes	Alternative topic/model approaches (LDA, logistic regression, Random Forest); pre- versus post-CDS implementation and alert-fired versus non-fired sessions
Chin-Purcell (2025)	Mixed-methods	Other; Error analysis	None / not stated	Yes	Human (researcher/clinician)
Creed (2022)	Mixed-methods protocol	Completed cross-validation and distinct held-out test-set evaluation of prototype models; planned preliminary system validation before trial deployment	User-centred design; usability testing; implementation-readiness assessment	Yes	Expert human ratings/manual coding; services-as-usual supervision and quality assurance
Dabravolskaj (2025)	Quantitative	Cross-validation	None / not stated	Yes	Human (researcher/clinician)
Goedken (2025)	Mixed-methods	Human validation of AI-generated outputs against the raw datasets; no quantified performance metrics	None / not stated	No	No formal comparator; trained qualitative researchers and content experts reviewed outputs
Golz (2022)	Qualitative	None / not stated	None / not stated	No	No comparator
Griffen (2024)	Quantitative	None / not stated	User satisfaction survey	Yes	Standard manual process
Ironside (2025)	Mixed-methods	None / not stated	Provider engagement analysis and post-implementation focus groups addressing acceptability, reluctance, time constraints, and workplace fit	No	A separate expert-human-feedback trial was described for contextual comparison, but there was no direct AI-versus-human comparator within the AI trial
Jones (2022)	Comparative case-study report	Planned precision/F-measure assessment; no completed AI evaluation reported	None / not stated	No	No comparator reported
Knippen (2025)	Mixed-methods	Qualitative triangulation and continuity with manual coding; no formal AI performance metrics	None / not stated	Yes	Manual reflexive thematic analysis and researcher-developed codebook
Lee (2026)	Quantitative	External validation	None / not stated	Yes	Human (researcher/clinician)

Li (2021)	Mixed-methods	No formal performance evaluation; boosted-tree sensitivity analysis of Random-Forest variable-importance findings	None / not stated	Yes	Another AI model (boosted trees) used as a sensitivity analysis
Lindamer (2022)	Mixed-methods	None / not stated	User satisfaction survey	No	No comparator
Mac Aonghusa (2020)	Quantitative methodological development	Internal performance assessment; error analysis; expert checking of predictions	None / not stated	Yes	Human expert annotations and review
Michie (2017)	Mixed-methods protocol	Planned inter-rater reliability; accuracy/precision/recall; prediction accuracy; comparison with published systematic reviews	Planned user evaluation of accuracy, salience, validity, and utility	Yes	Manual annotation; human experts; published systematic reviews / conventional methods
Moussa (2021)	Mixed-methods	Cross-validation	None / not stated	No	No comparator
Rappon (2021)	Quantitative	Cross-validation	None / not stated	No	No comparator
Rinne (2024)	Mixed-methods	Other	User satisfaction survey	Yes	Human (researcher/clinician)
Rosen (2025)	Quantitative	Reference-standard validation against 200 human-coded notes; sensitivity and specificity reported; no external validation	None / not stated	Yes	Human-coded psychotherapy notes
Rosen (2026)	Qualitative	None / not stated	User satisfaction survey	No	No comparator
Shankar (2025)	Mixed-methods	Topic coherence/perplexity; hybrid-classifier accuracy; manual sentiment-validation sample; comparison with fine-tuned healthcare BERT models	Stakeholder validation and co-design workshops	Yes	Manual review and alternative transformer models
Sibley (2025)	Mixed-methods	None / not stated (the RCT evaluated the implementation package rather than AI-model performance)	Acceptability, appropriateness, feasibility, satisfaction, credibility and perceived accuracy ratings; usage logs; supervisor interviews; end-of-study focus group	Yes	Standard STAND implementation supports without AI-generated MI feedback or the Care4 package
Stanfield (2024)	Quantitative	Cross-validation	None / not stated	Yes	Non-AI technology
Swanson (2025)	Mixed-methods	Iterative error detection and human validation; no quantified model-performance metrics	None / not stated	No	Human-coded database and study-team review were used for validation, not as a formal comparator
Wang (2016)	Quantitative	Error analysis	None / not stated	Yes	Another AI model
Ye (2025)	Quantitative	Internal-consistency/reliability assessment; stakeholder construct validation; SEM model-fit assessment; no train/test prediction validation	Practice-facilitator focus group reviewed the appropriateness, acceptability, and feasibility of selected features	No	No formal comparator
Zhong (2025)	Mixed-methods framework development	None / not stated (AI outputs were not independently evaluated)	None / not stated for the AI tool; stakeholder and expert validation concerned the CFGI	No	Human review and triangulation informed framework development but were not a formal AI comparator

Table A6. Outcome domains reported or prospectively specified (N = 31 application-focused sources).

Outcome domains coded for the 31 sources that described a specific AI system or prototype. Y = reported or prospectively specified; U = unclear; blank = not reported or specified. Domains: Time = efficiency/time outcomes; Cost = cost or resource outcomes; Acc = technical accuracy or performance; HF = human-centred outcomes (e.g., usability, satisfaction); Impl = implementation-specific outcomes (e.g., fidelity, stage, workflow); Eq = equity-related outcomes; Clin = clinical or health-system outcomes.

Source	Time	Cost	Acc	HF	Impl	Eq	Clin	Indicators reported (free text)
Bartholomew (2022)			Y		Y	U	Y	Model identified key indicators and produced a prioritized set of ZCTAs for SSP implementation; regression performance and variable importance reported.
Barwise (2026)	Y			U	Y	Y		Final trial: 44% of intervention admissions versus 41% of controls received an in-person interpreter (OR 1.40, 95% CI 0.98-2.00; p=0.061); time to first interpreter did not differ (HR 1.02, 95% CI 0.81-1.29). The companion interim abstract (Strechen 2024) reported 49% versus 31% and demonstrated feasibility during rollout. No patient-centred or cost outcomes were measured; stakeholder and nurse-perception evaluations were incomplete or planned.
Berkel (2023)			Y		Y	U		Concurrent validity with COACH ratings; predictive validity for engagement indicators; reported improvements in mean squared error over baseline for both fidelity and engagement prediction.
Chan (2025)	Y	Y	Y	Y	Y	Y	Y	AI extraction/classification performance (e.g., precision/recall/F1; qualitative coding agreement; semantic equivalence; hallucination/bias checks); user experience/usability; implementation outcomes + quintuple-aim KPIs and cost-effectiveness.
Chen (2025)	Y		Y		Y			The final study identified 14 EHR-activity topics. Random Forest distinguished after- from before-check-in sessions with accuracy 0.82, AUC-ROC 0.88, and weighted F1 0.83 on a held-out test set. Topic-derived prevalence and time measures identified shifts in CDS-focused, data-review, and documentation activities. A companion 2023 abstract reported an earlier 12-topic analysis, including 32-35-second increases in CDS-related work and compensatory decreases in other EHR activities.
Chin-Purcell (2025)	U	U	Y					Inter-coder reliability / agreement metrics (e.g., Krippendorff's alpha) and thematic similarity (free text).
Creed (2022)	Y	U	Y	Y	Y		Y	Completed prototype evaluation used cross-validation and a distinct 30% test set; most CTRS-code predictions exceeded 80% of human reliability and the total score reached 100% of human reliability. Planned outcomes include usability/readiness, acceptability, appropriateness, feasibility, therapist CBT fidelity, client PHQ-9/GAD-7 outcomes, dropout, service efficiency, and engagement. Cost efficiency is discussed but no formal cost outcome is specified.
Dabravolskaj (2025)	Y		Y		Y		Y	Matrices of 17 core functions and 55 forms of intervention activities and implementation strategies; ML was used to maximize efficient use of all available data and agreement with manual labels was used to assess classifier performance.
Goedken (2025)	U		U		Y			VA GPT(Beta) organized, categorized, identified themes, generated insights, and summarized data from 12 interviews and 20 post-pilot surveys. Outputs were checked against raw data and were described as timely and useful for leadership decisions, but efficiency and accuracy were not quantified.
Golz (2022)				Y	U			Described sentiments toward implemented IT using word frequencies and sentiment polarity; reported that many statements about technology were negative overall
Griffen (2024)			Y	Y	Y		Y	Therapist implementation fidelity (% correct steps); child independent correct handwashing responses (%); usability via System Usability Scale (SUS).
Ironside (2025)				Y	Y			In the AI-feedback trial, positive provider and supervisor attitudes toward MI predicted initial engagement (provider OR 3.3, 95% CI 1.0-10.7; supervisor OR 3.0, 95% CI 1.1-8.6); no measured factors predicted ongoing engagement. Focus-group themes included provider reluctance, time/productivity constraints, and structural workplace changes. The separate human-feedback trial was not a direct comparator.
Jones (2022)			Y		Y	U		Planned outcomes include algorithm precision/F-measure, identification of a virtual cohort, survey recruitment reach, and implementation-barrier assessment; no completed AI performance or implementation results were reported.

Knippen (2025)			U		Y			AILYZE-generated analyses were compared qualitatively with three rounds of manual reflexive thematic analysis. The authors reported continuity with manual coding and added analytic reflexivity but no quantitative AI performance, time, usability, or clinical outcome. The AI-assisted analysis contributed to six implementation themes and framework mapping.
Lee (2026)	Y	Y	Y		Y		Y	Trial 1: NLP identified 22,187 screen-positive passages (0.8%) from 2.6M EHR passages; reviewers adjudicated 7,494 passages over 34.3 abstractor-hours; 92.6% patient-level sensitivity. Trial 2: NLP identified 8,952 screen-positive passages (1.6%) from 559,596 passages at near-100% sensitivity; adjudicated 3,509 passages over 27.9 abstractor-hours. Efficiency gains vs exhaustive human abstraction demonstrated.
Li (2021)			U		Y		Y	Random-Forest and boosted-tree analyses examined the relative importance of transitional-care strategies, strategy groups, and covariates for 30-day readmission. No model-accuracy metrics were reported. Patient-reported receipt of strategies was more influential than hospital-reported delivery, while comorbidities were the strongest predictors.
Lindamer (2022)				Y	Y			Implementation outcomes across RE-AIM domains (reach/adoption/implementation/maintenance indicators) plus attendance and overall satisfaction; NLP used to summarize common free-text themes.
Mac Aonghusa (2020)	U	U	Y					Automated extraction sensitivities varied substantially by entity; outcome predictions were checked by behavioural scientists and described as reasonably accurate, but implementation or clinical effects were not evaluated.
Michie (2017)	Y		Y	Y	U			Planned evaluation includes time/efficiency relative to human methods, annotation and prediction accuracy, comparison with published systematic reviews, and user assessments of accuracy, salience, validity, and utility; downstream implementation effects were not yet specified.
Moussa (2021)			Y		Y		U	Random forest achieved 96.9% prediction accuracy. Outputs include predicted resolution rates (PRP) per strategy category across barriers (e.g., ‘empower stakeholders to develop objectives and solve problems’ had the highest predicted resolution across many barriers).
Rappon (2021)			Y		Y		U	Model performance reported with diagnostic odds ratio (DOR=5.2) and sensitivity/specificity (details in abstract).
Rinne (2024)	Y		U	Y	Y			Qualitative themes about EHRs and implementation science (strategies, outcomes, future EHR advancements); survey findings that ChatGPT provided efficient high-level overview but lacked nuance/context and required extensive editing; time saved was offset by revision work.
Rosen (2025)			Y		Y		Y	Rule-based NLP validation against 200 human-coded notes yielded sensitivity/specificity of 0.957/0.852 for prolonged exposure, 0.920/0.865 for cognitive processing therapy, and 0.931/0.859 for EMDR. These classifications served as trial implementation outcomes; PCL-5 symptom outcomes were also analyzed.
Rosen (2026)	Y		U	Y	Y			Overall positive experiences; concerns about prolonged response times; utilization seen as dependent on time availability and workplace digital infrastructure; no considerable differences between inpatient and outpatient. Inpatient PTs envisioned between-session use for personal knowledge enhancement; outpatient PTs envisioned in-session use for patient education.
Shankar (2025)	U		Y	Y	Y	Y	Y	The hybrid text classifier achieved 82.4% accuracy; manual review of 1000 sentiment classifications found 78.3% overall accuracy (86.2% for positive, 81.4% for negative, and 67.5% for ambiguous statements). Stakeholder workshops validated findings and co-developed five initiatives. Reported changes included an 18% increase in waiting-time satisfaction, 15% improvement in communication ratings, and 23% reduction in billing complaints. Efficiency was asserted but not directly measured.
Sibley (2025)	Y	U		Y	Y			The AI-assisted group completed STAND more efficiently and attended fewer appointments, but MI implementation quality declined over time relative to standard supports. The study assessed acceptability, appropriateness, feasibility, satisfaction, perceived feedback accuracy/credibility, usage, therapist knowledge/skills/motivation, and supervisor practices. Costs and client mental-health outcomes were not measured.

Stanfield (2024)	U		Y					Metamodel predictive performance vs simulation outputs (e.g., variance explained, absolute percent error); projected cancer cases and life-years lost across intervention scenarios; also evaluated fidelities of each model with R2.
Swanson (2025)			U		Y			ChatGPT initially miscounted consultation notes and required refinement prompts and team review to correct formatting-related and categorization errors. No quantified model-performance or efficiency outcome was reported. AI-assisted synthesis supported identification of consultation functions/forms, competency mapping, and gap analysis.
Wang (2016)	U		Y		Y			Primary outcomes relate to automated monitoring of implementation progress/stage (classification accuracy vs SIC); intended to reduce burden of manual coding for ongoing monitoring.
Ye (2025)			Y	Y	Y			Internal-consistency estimates for the five contextual-factor domains ranged from 0.71 to 0.89; practice facilitators reviewed feature appropriateness and construct alignment; SEM model fit was assessed using RMSEA. The analysis related five facilitation strategies to completed QI interventions and CPCQ change, but did not use train/test predictive validation.
Zhong (2025)								ChatGPT-4.0 classified candidate constructs and proposed category labels. Stakeholders and experts validated the resulting CFGI, not the AI tool; no AI-specific performance, human-centred AI, or implementation outcome was reported.

Table A7. Risks, adverse effects and unintended consequences (N = 40).

Risk categories coded for all 40 sources. Y = explicitly discussed; U = unclear; blank = not discussed. AI-specific = bias/fairness, hallucinations, prediction error, opacity; Workflow = workflow disruption, poor fit with service delivery; User = overreliance, trust, acceptability, de-skilling; Clinical = clinical or patient-level harms.

Source	Risks reported	AI-specific	Workflow	User	Clinical	Details (free text)
Bartholomew (2022)	Yes	Y	U			Discusses black-box opacity, class imbalance, limited generalisability, differing accuracy across training and validation subsets, and dependence on data not routinely available; no empirical harm was reported.
Barwise (2026)	Yes	U	Y	U	U	Reports occasional data-pipeline failures, dependence on local interpreter and information-technology capacity, inaccurate or incomplete EHR language data, limited availability for lower-diffusion languages, a complexity score that omits social complexity, uncertain generalisability, and lack of patient-centred outcome and patient-voice assessment. No harm caused by the AI system was demonstrated.
Berkel (2019)	Yes	U	Y	Y		Discusses recording and secure-data infrastructure, privacy, cost and return on investment, trust and acceptability, limits of automated ratings, and risks of worsening inequities or excluding non-English speakers.
Berkel (2023)	No					
Brown (2013)	No					
Chan (2025)	Yes	Y	Y	Y	U	Risks discussed include data security/privacy, accuracy issues, hallucinations, bias, and user trust/acceptability; mitigation includes local hosting, governance, and evaluation safeguards.
Chen (2025)	Yes	Y	U		U	Topic labels required subjective domain-expert interpretation; model-derived time estimates were probabilistic and not validated against direct time measurement; unobserved confounding prevented causal attribution; audit logs omitted work outside the EHR; and transferability to other settings remained uncertain. The study also identified potential CDS-related workflow burden and patient-safety concerns, although these were not harms caused by the monitoring model.
Chin-Purcell (2025)	No					
Creed (2022)	Yes	U	Y	U		Anticipates technology-access and integration problems, provider turnover, client no-shows, missing data, and recruitment delays. It describes privacy and security risks for session recordings and mitigation through HIPAA-compliant cloud infrastructure, encryption, access controls, and de-identification. No empirical harm was reported.
Dabravolskaj (2025)	No					
El-Bassel (2025)	Yes	Y	U	U		Discusses potential AI risks (e.g., bias/fairness, other AI-specific risks) conceptually; no empirical harms reported.
Goedken (2025)	Yes	Y	U			States that trained qualitative researchers and content experts remain necessary to supply context, detect nuance, and ensure AI-output accuracy. No quantified error rate or empirical adverse effect was reported.
Golz (2022)	No					
Griffen (2024)	No					
Huguet (2024)	Yes	Y	U	U		Highlights risks/challenges including prediction bias, need for recalibration as context changes, and other methodological concerns when using ML to support implementation decisions.
Ironside (2025)	No					
Jones (2022)	Yes	U	Y	U		Discusses unequal social-media representation, privacy, algorithmic generalisability, changing technology platforms, and the need for advanced technical expertise; no empirical AI-related harm was reported.
King (2023)	Yes	Y	Y	U	U	Highlights potential risks that data science efforts could ingrain existing biases, widen health disparities, or disrupt clinical workflows.
Knippen (2025)	No					

Lee (2026)	Yes	Y	Y		U	Discusses risk of NLP misclassification (screen threshold tuned for near-100% sensitivity to mitigate false negatives); acknowledges that pragmatic trials require low measurement error, motivating hybrid human-in-the-loop design.
Li (2021)	No					
Lindamer (2022)	No					
Mac Aonghusa (2020)	Yes	Y	U	Y	U	Reports variable extraction sensitivity and challenges generalising algorithms to heterogeneous real-world reports; discusses opacity, fairness, bias, accountability, transparency, and appropriate user trust in predictions.
Michie (2017)	Yes	Y		Y		Anticipates concerns about the black-box nature, reliability and trustworthiness of predictions, and the need to communicate uncertainty and support appropriate user trust; no empirical harms were reported.
Moussa (2021)	No					
Pearson (2020)	Unclear					Discusses need to define harms and reduce harm/waste as part of evaluation/implementation research (not empirically measured).
Rappon (2021)	No					
Rinne (2024)	Yes	Y		Y		LLM lacked nuance and contextual understanding; required extensive human editing; concerns about erroneous citations (avoided by instructing 'do not include references'); authors note need for close supervision and attentive human involvement.
Rosen (2025)	No					
Rosen (2026)	Yes	Y	Y	Y	U	Slow response times discouraged in-session use; workplace digitization gaps; concerns about validity/trustworthiness of LLM-generated summaries; ethical considerations about AI's role in clinical decision-making and patient care flagged for future research.
Shafran (2026)	Yes	Y	Y	Y	Y	Details risks of AI in CBT training: data quality/bias, noise in ML training data, misclassification of competence by rigid fidelity rules (e.g., suicide risk), privacy of sensitive clinical materials, data ownership/consent ambiguity, copyrighted training data issues, cultural/linguistic misalignment in low-resource settings, environmental/carbon footprint, overreliance on AI judgement, and need for human oversight and explainability.
Shankar (2025)	Yes	Y	U	U		Reports limited handling of sarcasm, negation, ambiguous language, and domain terminology; self-selection and cultural bias in feedback data; limited cross-context generalisability; and organizational integration barriers. Transformer models improved accuracy but introduced higher computational cost and black-box interpretability challenges that could impede stakeholder engagement.
Sibley (2025)	Yes		Y	Y	U	AI-assisted supports were associated with declining MI quality and appeared to foster a task-oriented therapist mindset that prioritized efficiency over relational process. Therapists also reported navigation and workflow difficulties. Potential consequences for client engagement or outcomes were raised but not measured.
Stanfield (2024)	No					
Swanson (2025)	Yes	Y	U			ChatGPT miscounted notes and required clarification to avoid formatting-related and category errors. The paper discusses hallucination, response degradation, oversimplification, contextual misinterpretation, and inaccurate AI-generated notes or transcripts; all outputs required human validation.
Trinkley (2024)	Yes	Y	Y	Y	U	Discusses cautions/pitfalls for AI in implementation science (e.g., bias/fairness, transparency/interpretability, workflow disruption, overreliance/trust issues, monitor unintended consequences).
Wang (2016)	No					
Ye (2025)	Yes	Y				Acknowledges the risk of spurious correlations, limited robustness, unmeasured confounding, reverse causation, selection bias from excluded practices, and limited geographic/system generalisability. The authors plan application to other implementation datasets for validation.

Younas (2024)	Unclear	U	U	U	U	Mentions the promise of AI to improve precision/accuracy across implementation stages; specific unintended consequences/harms not clearly detailed in available text.
Zhong (2025)	No					