Review Article

When Does Human–AI Collaboration Improve Clinical Pharmacy? A Realist Review of Trust, Workload, and Decision Quality

George Wilson[1]*, Chloe Bennett[2], Ethan Wright[1], Jack Turner[3]

[1]Department of Human-AI Collaboration in Pharmacy, Faculty of Pharmacy, University of Dundee, Dundee, United Kingdom. [2]Department of Trust and Workload in Clinical Pharmacy, Faculty of Pharmacy, Newcastle University, Newcastle, United Kingdom. [3]Department of Decision Quality and AI Partnership, Faculty of Pharmacy, University of Aberdeen, Aberdeen, United Kingdom.

Abstract

Human–AI collaboration in clinical pharmacy is often evaluated through average accuracy or efficiency, although the same system may reduce cognitive burden in one setting while increasing interruption, misplaced trust, or responsibility ambiguity in another. This realist review examined when, how, for whom, and under what conditions human–AI collaboration may improve medication-related decision quality, workload, and safety. We used a RAMESES-aligned realist-review design. Searches covered seven databases and supplementary citation, journal, web, and registry sources. After deduplication, 817 records were screened, 79 reports were sought, 76 full texts were assessed, and 21 reports representing 21 distinct study or document families were included. Evidence was selected for relevance and rigour, extracted into context–mechanism–outcome configurations, tested against contradictory cases, and synthesized into a refined programme theory. Collaboration appeared most promising when AI supplied complementary, task-specific information; pharmacists retained interpretive authority; outputs were integrated into workflow; and escalation routes remained explicit. Workload reduction could preserve cognitive capacity when low-value processing was removed, but poorly targeted alerts redistributed rather than reduced work. Trust was more appropriately calibrated when system limits, uncertainty, and accountability were visible. Incorrect recommendations, opaque explanations, weak organizational support, and unclear responsibility could promote automation bias, vigilance loss, or decision diffusion. workflow, professional authority, organizational support, and system reliability. The resulting programme theory is evidence-bounded and requires prospective testing in clinical-pharmacy settings before routine implementation.

Keywords: Realist review, Digital pharmacy, Artificial intelligence, Pharmacy practice, Medication safety, Evidence synthesis

Introduction

Human–AI collaboration is increasingly framed as a route to safer, faster, and more consistent clinical work. Yet collaboration is not a single intervention. It may involve an alert, ranked recommendation, predictive score, conversational assistant, or decision-support interface, and each configuration redistributes attention, interpretation, and authority differently. The broader vision of high-performance medicine emphasizes complementarity between computational pattern recognition and human contextual judgment rather than substitution of one by the other [1].

In clinical pharmacy, however, medication decisions combine technical evidence with patient goals, comorbidity, workflow constraints, uncertainty, and professional accountability. An AI output may therefore improve one component of a decision while simultaneously creating new work, suppressing doubt, or obscuring who must act.

Address for correspondence: George Wilson, Department of Human-AI Collaboration in Pharmacy, Faculty of Pharmacy, University of Dundee, Dundee, United Kingdom.
g.wilson@dundee.ac.uk
Received: 10 April 2026; **Accepted:** 21 August 2026

How to cite this article: Wilson G, Bennett C, Wright E, Turner J. When Does Human–AI Collaboration Improve Clinical Pharmacy? A Realist Review of Trust, Workload, and Decision Quality. Arch Pharm Pract. 2026;17(3):66-74. https://doi.org/10.51847/hvKJBM7bs8

Empirical studies outside pharmacy illustrate this variability. In skin-cancer recognition, human–computer collaboration altered diagnostic performance, but benefit depended on the interaction between clinician and system rather than on model accuracy alone [2]. A deep-learning assistant used for liver-cancer histopathology likewise demonstrated that assistance can change classification behaviour and error patterns, not merely add an independent prediction [3]. These findings are relevant to medication work because pharmacists also integrate algorithmic suggestions with incomplete records, competing risks, and time-sensitive judgments. They do not establish pharmacy effectiveness, but they show why an average comparison between unaided and AI-assisted performance can conceal consequential differences among users, tasks, and settings.

A recent synthesis of human–large language model collaboration in clinical medicine further indicates that reported effects vary across study designs, specialties, tasks, and outcome definitions [4]. Average-effect questions such as "Does AI improve decisions?" are therefore insufficient. They compress heterogeneous contexts and mechanisms into a single estimate and may treat workload reduction, speed, confidence, agreement, and clinical quality as interchangeable. Realist inquiry instead asks what resources an intervention offers, how participants reason or respond to those resources, and which outcomes follow under particular conditions.

This review accordingly examined six linked questions: when complementary expertise improves medication-related decisions; when workload relief preserves cognitive capacity rather than reducing vigilance; how interruption and alert burden reshape collaboration; what supports calibrated trust; when automation bias or responsibility diffusion emerges; and how organizational support and escalation pathways modify outcomes. The initial programme theory proposed that improvement is most likely when AI addresses a bounded task, provides information that complements rather than duplicates professional expertise, fits the medication-use workflow, communicates limitations, and leaves interpretive authority and escalation responsibility explicit. The review tested, contradicted, and refined that theory rather than assuming collaboration was inherently beneficial. Its purpose was explanatory rather than prescriptive: to identify recurring context–mechanism–outcome patterns, preserve uncertainty where evidence conflicted, and distinguish changes in model or task performance from demonstrated improvements in medication safety, patient outcomes, implementation durability, or professional practice overall.

Review Design, Questions, and Protocol

This realist review examined how human–AI outcomes arise from interactions among context, system resources, professional reasoning, and organizational conditions. Sources were assessed for relevance to programme theory and rigour of supporting evidence [5, 6]. The review sought explanatory context–mechanism–outcome patterns rather than a pooled effect estimate.

The scope covered AI-assisted medication review, prescribing, clinical pharmacy, medication safety, and related tasks in which professionals interpreted or acted on algorithmic outputs. Purely technical prediction studies were excluded. Outcomes included decision quality, workload, trust, reliance, escalation, responsibility, and implementation effects.

Initial programme theory proposed that bounded, workflow-integrated AI may improve decisions when it supplies relevant information while preserving professional authority, but may cause harm through interruption, overreliance, or unclear accountability [7]. No formal stakeholder panel was conducted. Reporting followed RAMESES principles, with PRISMA 2020 used only for the selection diagram [8].

Eligibility, Searching, and Extraction

Eligible reports published from 2017 to 2026 examined human–AI interaction in medication-related work or mechanisms such as trust, workload, attention, responsibility, and decision quality. Experimental, implementation, qualitative, mixed-methods, simulation, and relevant review evidence was included. Generic clinical evidence was retained only when transferability to medication work could be justified.

Seven databases and supplementary citation, journal, web, and registry searches were used. Records, reports, and studies were distinguished according to PRISMA guidance [9, 10]. Deduplication and screening remained subject to accountable human judgment despite the possible use of machine-assisted tools [11]. Multi-source searching addressed differences in database coverage [12].

Extraction covered setting, role, task, AI function, workflow position, autonomy, uncertainty, training, organizational support, outcomes, maturity, and limitations. Trust, workload, and decision quality were defined broadly and were not reduced to agreement, speed, or technical accuracy.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for eligibility, definitions, and evidence-extraction framework for the review When Does Human–AI Collaboration Improve Clinical Pharmacy? A Realist Review of Trust, Workload, and Decision Quality.

Table 1. Eligibility, definitions, and evidence-extraction framework for the review When Does Human–AI Collaboration Improve Clinical Pharmacy? A Realist Review of Trust, Workload, and Decision Quality.

Review Element	Operational Definition	Eligibility or Decision Rule	Data Field	Rationale	Bias or Ambiguity Risk	Planned Synthesis Use
Human–AI Collaboration	Observable human receipt, interpretation, use, rejection, questioning, or escalation of an AI-generated output	Include only when the human contribution or response can be identified; exclude standalone model-performance studies	Human role; AI role; interaction point; final decision authority	Realist explanation requires analysis of how intervention resources are interpreted and enacted [5]	"Collaboration" may be inferred merely because a clinician is present	Distinguish augmentation, advisory support, verification, and de facto automation
Clinical-Pharmacy Relevance	Direct medication-related work or a transferable clinical task involving comparable uncertainty, risk, and professional judgment	Include direct pharmacy evidence; retain adjacent clinical evidence only when transferability and limits can be stated	Setting; medication task; professional group; transferability judgment	Review-derived eligibility boundary intended to prevent broad clinical-AI evidence from being treated as pharmacy evidence	Transferability may be overstated when workflow and authority differ	Test whether mechanisms recur across direct and adjacent settings
Context	Pre-existing condition that enables, constrains, or modifies the response to an AI resource	Extract only conditions relevant to the proposed mechanism or outcome	Task complexity; staffing; integration; training; reliability; authority; organizational support	Context is analytically distinct from the intervention and mechanism in realist synthesis [5]	Descriptive setting characteristics may be mislabeled as causal contexts	Compare configurations across enabling and constraining conditions
Mechanism	Change in reasoning, attention, confidence, motivation, or action triggered by an intervention resource	Do not label a technology, feature, or observed outcome as a mechanism	Resource offered; participant response; inferred reasoning; evidential basis	Mechanism claims require relevance, evidential richness, and sufficient rigour for the intended inference [6]	Mechanisms may be speculative or reconstructed from limited reporting	Develop and test complementary expertise, cognitive relief, interruption, trust, bias, and responsibility configurations
Outcome	Observable consequence of the interaction, including decision, workload, safety, behavioural, or implementation effects	Retain reported outcomes; do not infer clinical benefit from model accuracy, speed, agreement, or confidence alone	Outcome definition; denominator; direction; timing; measurement method	Review-derived rule separating technical performance from clinical usefulness	Surrogate outcomes may be interpreted as patient benefit	Map proximal and distal outcomes while preserving evidential distance
Evidence Relevance And Rigour	Relevance is contribution to programme theory; rigour is credibility of the evidence used for a particular inference	Appraise each source according to the specific claim it supports rather than assigning one universal score	Supported inference; design strength; reporting adequacy; limitation; confidence	Realist appraisal is purpose-dependent and should examine relevance, richness, and rigour [6]	A methodologically strong study may be irrelevant, while a weaker study may contain useful explanatory detail	Weight configurations and define interpretive boundaries
Iterative Theory Testing	Repeated comparison of candidate explanations with supportive, contradictory, and rival evidence	Retain negative cases and revise propositions when configurations fail across contexts	Initial proposition; supportive case; contradictory case; revision; unresolved uncertainty	Realist methods permit iterative movement between evidence and theory refinement [7]	Selective use of confirming evidence may make the theory unfalsifiable	Refine contingent propositions rather than produce an average-effect conclusion
Records, Reports, And Studies	Records are database entries, reports are retrieved documents, and studies or document families are underlying investigations	Preserve separate identifiers and link multiple reports from one study family	Record ID; report ID; study-family ID; duplicate-family ID	Transparent review accounting requires separation of identification and inclusion units [9]	Conflation can inflate evidence counts or obscure companion reports	Reconcile the flow diagram and evidence-base classifications
Heterogeneous Evidence	Empirical and methodological sources contributing different forms of evidence to programme theory	Chart source type and contribution without treating all sources as equivalent	Design; setting; contribution; validation level; synthesis role	Heterogeneous evidence requires explicit classification and transparent charting [10]	Narrative synthesis may flatten differences in design or evidential strength	Compare mechanism support while preserving source-specific limitations
Screening Support	Computational assistance used only to organize or prioritize records while human reviewers retain decisions	Do not permit automated prioritization to determine final eligibility independently	Tool use; prioritization method; human verification; audit trail	Automation may improve review efficiency but requires transparent and accountable human oversight [11]	Low-ranked relevant evidence may be missed if prioritization is treated as exclusion	Document the boundary between assistance and review authority

Information-Source Coverage	Use of multiple bibliographic and supplementary sources to reduce dependence on one search system	Search all prespecified sources and retain source-specific counts	Database; platform; query; filters; search date; records captured	Academic search systems differ in coverage, precision, and reproducibility [12]	Uneven indexing, inaccessible records, and database overlap may distort retrieval	Interpret evidence gaps in relation to search coverage and indexing limitations

Search, Selection, Extraction, Appraisal, and Synthesis

Search and Selection Process

Selection proceeded through deduplication, title and abstract screening, full-text retrieval, and eligibility assessment. Uncertain records advanced to full text, and disputed decisions were resolved through documented discussion. Procedures used prespecified questions, explicit decisions, and reproducible documentation, consistent with guidance for medical evidence synthesis [13]. At full text, each excluded report received one primary reason from a mutually exclusive hierarchy. Reports from the same study were linked, and background or methodology sources were distinguished from reports contributing to programme theory.

CMO Construction and Contradictory-Case Analysis

For each included report, extracted observations were coded as context, intervention resource, mechanism, and outcome. Mechanisms were treated as changes in reasoning, attention, confidence, motivation, or action triggered by available resources, not as labels for the technology itself. Candidate configurations were compared across pharmacy and transferable clinical settings. Findings were examined for explanatory fit, methodological credibility, and applicability to medication work. Contradictory cases were retained to test whether beneficial mechanisms reversed under different task demands, reliability levels, workflow arrangements, or accountability structures.

Theory Testing, Refinement, and Reporting

Theory testing used iterative comparison among the initial propositions, supportive cases, negative cases, and alternative explanations. The programme theory was refined only when evidence clarified how a context enabled or constrained a mechanism and how that mechanism was linked to an outcome. Reporting decisions followed updated systematic-review guidance on transparent methods and selection accounting [14]. Claims comparing AI with clinicians were interpreted cautiously because reviews of clinical deep-learning studies have identified weaknesses in design, reporting, and claims of superiority [15]. No pooled effect estimate was planned because the included evidence differed in intervention, task, setting, comparator, and outcome. Synthesis therefore emphasized recurring configurations, contradictions, evidence boundaries, and propositions requiring prospective clinical-pharmacy evaluation under routine conditions and across organizational environments.

Search and Selection Results

The database and supplementary searches identified 1,134 records: 1,048 from seven databases and 86 from citation searching, target-journal handsearching, and focused web or registry sources. Deduplication removed 317 records, leaving 817 for title and abstract screening. Of these, 738 were excluded and 79 reports were sought. Three reports could not be retrieved, so 76 full texts were assessed. Fifty-five were excluded: 18 evaluated an algorithm without human–AI collaboration; 12 lacked clinical-pharmacy or transferable medication-decision relevance; nine could not inform a context–mechanism–outcome explanation; six were ineligible publication types; four fell outside the date window; three were duplicate or companion reports without additional relevant evidence; and three lacked sufficient full-text detail or a stable identifier. Twenty-one reports representing 21 distinct study or document families entered the realist synthesis.

Figure 1 presents the completed review-selection pathway in a black-and-white RAMESES-compatible PRISMA-style flow diagram for evidence selection, documenting identification, deduplication, screening, eligibility, exclusions with reasons, and final inclusion using the approved record-accounting values.

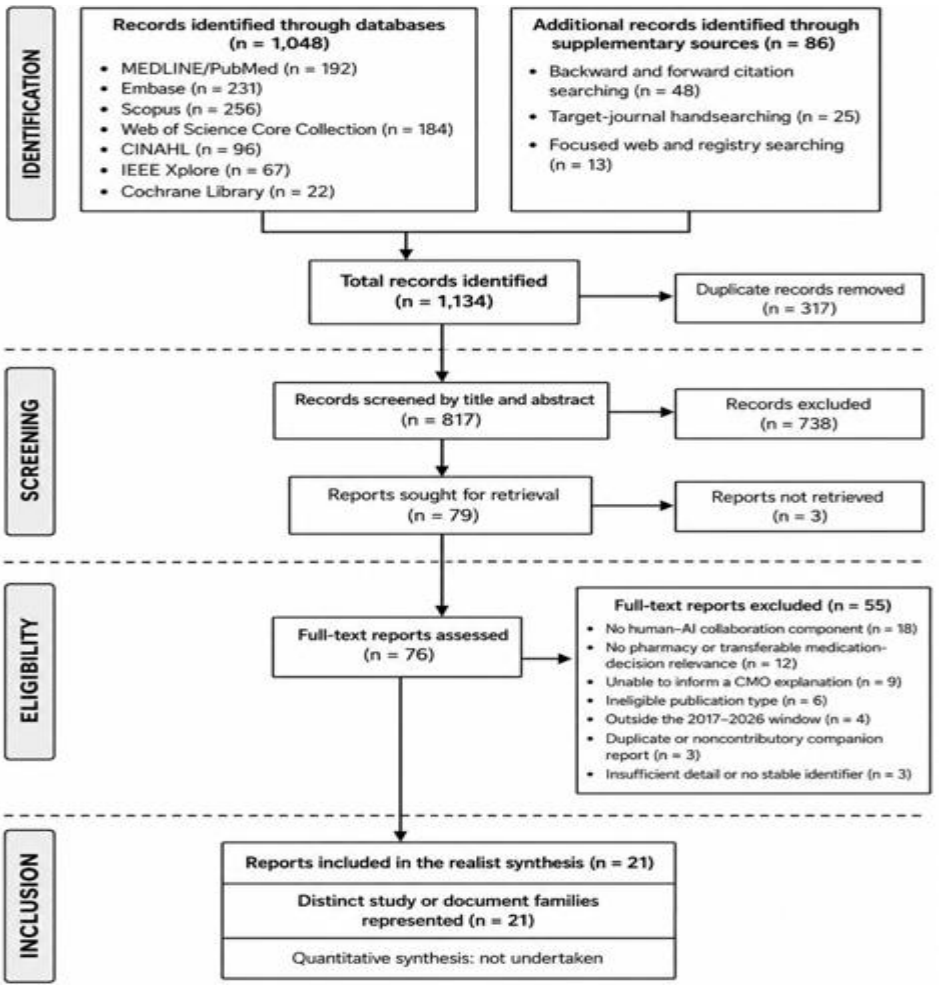

Figure 1. Black-and-White RAMESES-Compatible PRISMA-Style Flow Diagram for Evidence Selection

Evidence-Base Characteristics and Initial Programme Theory

The included evidence was heterogeneous in design, setting, intervention maturity, and outcome definition. A scoping review of machine-learning approaches to medication alerts described attempts to improve the specificity and usefulness of clinical decision support, while also showing that prospective evaluation and workflow evidence remained limited [16]. A systematic review of trust in AI-based clinical decision support identified clinician, system, task, and organizational factors that influence willingness to rely on recommendations [17]. A broader scoping review of AI-assisted medical decision-making similarly found that effectiveness depends on the interaction among recommendation quality, user expertise, task characteristics, presentation, and implementation context [18]. These sources supported a programme theory based on contingent interaction rather than uniform technological effect.

The initial programme theory proposed four linked conditions. First, the AI resource must contribute information that is relevant to a bounded medication task. Second, the pharmacist must be able to interpret, question, and override the output. Third, workflow design must prevent assistance from becoming unmanaged interruption or hidden verification work. Fourth, the organization must sustain training, escalation, monitoring, and accountability. Under these conditions, complementary expertise and cognitive relief could improve decision quality. When one or more conditions were absent, the same system could activate interruption, overreliance, vigilance loss, or responsibility diffusion.

Evidence-Base Characteristics

CMO: Complementary Expertise

Pharmacy evidence showed that AI value depends on how professionals interpret and integrate recommendations, not merely on access to them [19]. Trials similarly found that AI may fail to improve reasoning when interaction is poorly optimized, but can support uncertain decisions when users critically assess its recommendations [20, 21]. Benefit therefore arises when AI supplies task-relevant information while pharmacists retain authority and compare outputs with patient-specific evidence.

CMO: Workload Relief and Cognitive Capacity

AI may reduce workload by prioritizing cases, screening records, or summarizing information. However, benefits disappear when verification, navigation, or documentation demands exceed the effort saved. Cognitive capacity improves only when low-value work is removed without weakening professional awareness or judgment.

CMO: Interruption and Alert Burden

Frequent, poorly timed alerts can cause interruption, habituation, and reduced responsiveness [22]. Role-sensitive design and more precise targeting may reduce burden [23, 24]. Effective alerts must remain relevant, interpretable, actionable, and linked to authority to intervene.

CMO: Calibrated Trust

Clinicians may follow incorrect AI advice, while explanations can increase confidence without ensuring correctness [25–27]. Calibrated trust therefore depends on visible reliability, uncertainty, task fit, and opportunities for independent checking.

CMO: Automation Bias and Responsibility Diffusion

Automation bias occurs when AI narrows reasoning or is accepted because of perceived authority. Responsibility diffusion occurs when verification, override, communication, and follow-up duties are unclear. Both excessive and insufficient reliance can weaken collaboration.

CMO: Organizational Support and Escalation

Trust, liability, ownership, monitoring, integration, and maintenance remain organizational concerns [28, 29]. Clear roles, training, workflow integration, escalation thresholds, fallback procedures, and continuous evaluation help convert uncertainty into appropriate review rather than silent acceptance or rejection [30].

Refined Programme Theory

The refined programme theory therefore treated outcomes as products of interacting conditions. Bounded tasks, complementary information, visible uncertainty, usable interfaces, professional authority, workload-sensitive integration, and supported escalation can activate comparison, cognitive relief, calibrated reliance, and accountable action. In contrast, low-specificity output, repeated interruption, opaque confidence, weak monitoring, and ambiguous ownership can activate habituation, automation bias, vigilance loss, or responsibility diffusion. These propositions describe evidence-supported and review-derived relationships, not measured effect sizes or universal causal rules.

Methodological appraisal qualified every configuration. Direct pharmacy evidence was concentrated in medication-review and alerting contexts, whereas several trust, diagnostic-reasoning, and automation-bias mechanisms came from adjacent clinical settings. Randomized studies strengthened inference about short-term assisted performance but often used bounded tasks or outcomes distant from routine medication safety. Reviews broadened coverage but inherited variation in terminology and measurement. Qualitative evidence offered richer explanations of workflow and responsibility, although causal direction was less secure. Across designs, accuracy, agreement, confidence, response time, and alert acceptance were not interchangeable with appropriateness, prevented harm, or patient benefit. Few sources followed collaboration long enough to establish adaptation, sustained workload change, maintenance effects, or consequences after system updates. Evidence concerning equity, patient involvement, interprofessional negotiation, and organizational escalation was sparse. The synthesis consequently assigned confidence to recurring mechanisms supported across designs, while treating setting-specific outcomes and cross-domain transfer as provisional. No quantitative synthesis was undertaken, and no certainty grade was converted into a claim of clinical effectiveness, because interventions, comparators, tasks, and outcome definitions were too heterogeneous for a defensible common estimate.

Discussion

When Collaboration Improves Decisions

The principal finding overall is that human–AI collaboration improves decision work only when technology, user, task, workflow, and organization form a compatible configuration. The review did not identify a general mechanism by which adding AI necessarily produces better clinical-pharmacy decisions. Benefit arose when a bounded system resource complemented professional knowledge, enabled comparison or prioritization, and remained open to questioning. Harmful or null outcomes arose when output quality was unreliable, task boundaries were unclear, interruptions accumulated, or organizational arrangements encouraged reliance without preserving accountability.

This interpretation aligns with responsible machine-learning guidance emphasizing that healthcare AI must be evaluated within clinical systems rather than as an isolated model [31]. In pharmacy, the relevant unit is the medication decision pathway: information is generated, interpreted, reconciled with patient circumstances, communicated, acted upon, and monitored. Each transition can amplify or lose benefit. A technically accurate output may arrive too late, duplicate known information, require disproportionate verification, or conflict with a pharmacist's authority. Conversely, a modest system may be useful if it reliably surfaces overlooked information when professional judgment can act.

Complementary expertise was conditional on difference with relevance. AI assistance added value when it processed information at a scale or speed that supported a defined human judgment. Mere disagreement was not evidence that the machine was informative, and agreement was not evidence that the combined decision was correct. The pharmacist required knowledge, time, and transparency to compare the recommendation with competing evidence.

Without those resources, the system could narrow reasoning rather than extend it.

When Workload Savings Produce New Risks

Workload findings similarly require separation of task removal from task redistribution. Screening, prioritization, or summarization may reduce visible processing time while creating less visible work through checking, documentation, alert management, exception handling, and follow-up. A workload-saving claim is credible only when evaluation captures the pathway and observes whether vigilance, communication, or unresolved queues deteriorate. Adaptation can also change effects over time: users may initially inspect every output, later develop efficient routines, or become habituated to low-value recommendations.

Explanation should not be treated as a universal remedy. Critiques of current explainable-AI approaches warn that understandable outputs may not establish fidelity, usefulness, or safety [32]. Explanations were potentially valuable when they helped users identify evidence, limits, and reasons for disagreement. They were potentially harmful when persuasive presentation increased confidence without improving the recommendation. Trust calibration therefore requires observable performance, task-specific uncertainty, independent checking, and feedback about errors, not explanation alone.

Prospective evidence also showed that access to advanced generative assistance does not ensure improved task performance. A randomized trial of GPT-4 assistance found gains on some patient-care tasks, but its controlled conditions and measured outcomes do not establish broad clinical-pharmacy benefit [33]. This reinforces the distinction between capability and collaboration. Evaluation should examine who uses the tool, for which decision, under what workload, with what fallback, and with which consequences when advice is wrong, ignored, or contested.

Figure 2 presents the original conceptual synthesis for refined programme theory of human–AI collaboration in clinical pharmacy, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

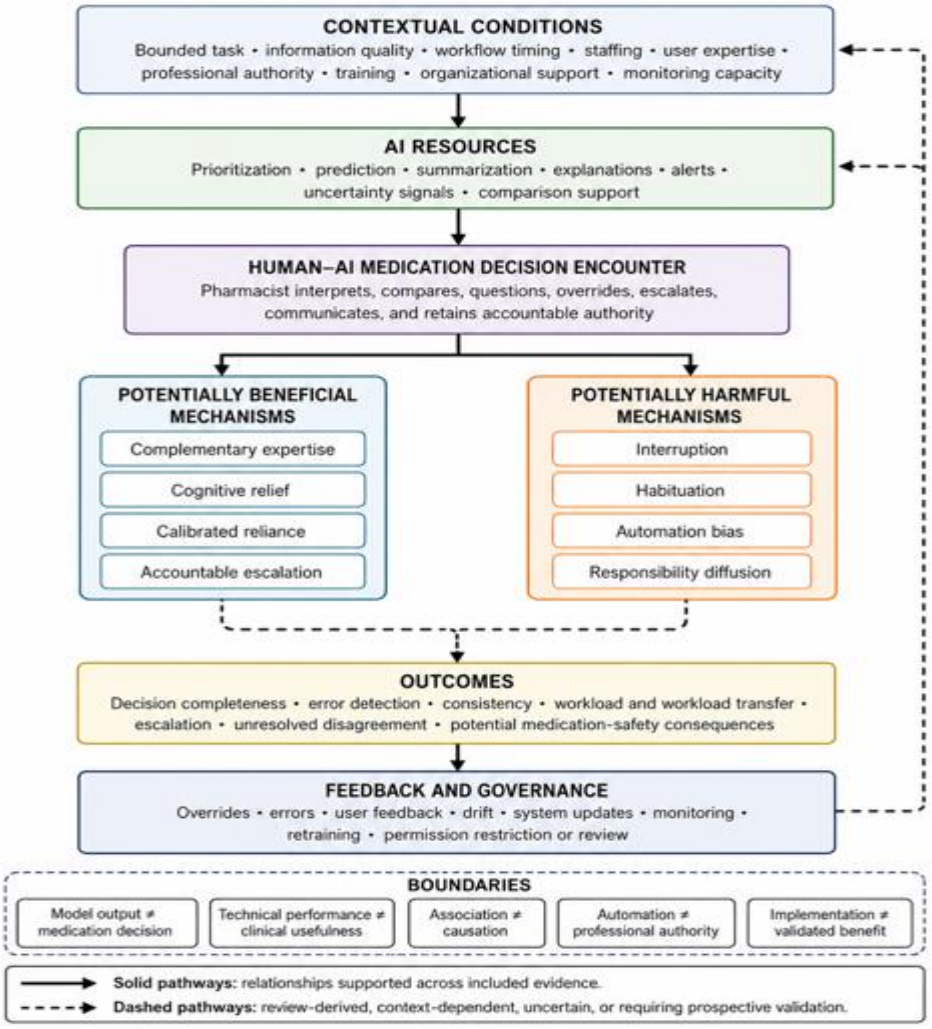

Figure 2. Refined Programme Theory of Human–AI Collaboration in Clinical Pharmacy

Refined Programme Theory and Implementation

Context—including task scope, data quality, workflow, expertise, authority, training, support, and monitoring—shapes how users respond to AI resources. Responses may include comparison, cognitive relief, calibrated reliance, interruption, automation bias, or responsibility diffusion, affecting decision quality, workload, escalation, and medication safety. Feedback from errors and overrides should inform system governance.

Local evaluation should test representative users and cases, including uncertain or incorrect outputs. Measures should separate model performance from human–AI performance and assess workload transfer, disagreement, escalation, and changing effects across settings and updates.

Strengths and Limitations

Human factors, workflow, safety, input handling, and real-world performance should be prospectively evaluated [34–36]. These frameworks support testing but do not validate the programme theory.

Strengths included mechanism-focused synthesis and attention to contradictory evidence. Limitations included heterogeneous evidence, few pharmacy-specific studies, reliance on short-term outcomes, and limited organizational, equity, patient, and longitudinal evidence. The propositions are therefore testable explanations rather than universal conclusions.

Conclusion

Human–AI collaboration in clinical pharmacy is best understood as a contingent relationship rather than a uniformly beneficial intervention. Across the included evidence, improved decision work was most plausible when AI addressed a bounded task, contributed information that complemented professional expertise, fitted the workflow, communicated uncertainty, and preserved the pharmacist's authority to question, override, and escalate. Under those conditions, comparison, prioritization, and cognitive relief could support more complete or consistent reasoning.

The same technology could produce different outcomes when these conditions changed. Low-specificity alerts, repeated interruption, opaque recommendations, unreliable output, time pressure, and ambiguous responsibility could redistribute workload, narrow reasoning, weaken vigilance, or promote inappropriate reliance. Explanations alone did not ensure calibrated trust, and technical performance did not establish medication-safety benefit, patient benefit, implementation durability, or organizational readiness.

The refined programme theory links task and organizational contexts to mechanisms of complementary expertise, cognitive relief, interruption, habituation, calibrated reliance, automation bias, and responsibility diffusion. It provides testable propositions for evaluating collaboration across medication-review, alerting, prescribing, and related decision pathways. It is not a validated predictive model or a clinical implementation standard.

Future evaluations should examine representative users, routine workflows, incorrect and uncertain outputs, workload transfer, overrides, unresolved disagreement, escalation, and downstream medication consequences over time. Until such evidence accumulates, health systems should interpret claims about collaborative AI cautiously. The central question is not whether humans or AI perform better in isolation, but whether their interaction preserves accountable professional judgment while producing outcomes that matter to patients and medication safety across clinical settings.

ACKNOWLEDGMENTS: None
CONFLICT OF INTEREST: None
FINANCIAL SUPPORT: None
ETHICS STATEMENT: None

References

1. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
2. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229-34. doi:10.1038/s41591-020-0942-0
3. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med. 2020;3(1):23. doi:10.1038/s41746-020-0232-8
4. Wang G, Zhang K, Jiang J, Wang C, Bi H, Liang H, et al. Human-large language model collaboration in clinical medicine: A systematic review and meta-analysis. NPJ Digit Med. 2026;9(1):195. doi:10.1038/s41746-026-02382-2
5. Jagosh J. Realist synthesis for public health: Building an ontologically deep understanding of how programs work, for whom, and in which contexts. Annu Rev Public Health. 2019;40:361-72. doi:10.1146/annurev-publhealth-031816-044451
6. Dada S, Dalkin S, Gilmore B, Hunter R, Mukumbang FC. Applying and reporting relevance, richness and rigour in realist evidence appraisals: Advancing key concepts in realist reviews. Res Synth Methods. 2023;14(3):504-14. doi:10.1002/jrsm.1630
7. Renmans D, Castellano Pleguezuelo V. Methods in realist evaluation: A mapping review. Eval Program Plann. 2023;97:102209. doi:10.1016/j.evalprogplan.2022.102209
8. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71
9. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160. doi:10.1136/bmj.n160
10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850
11. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3(2):125-33. doi:10.1038/s42256-020-00287-7
12. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods. 2020;11(2):181-217. doi:10.1002/jrsm.1378

13. Muka T, Glisic M, Milic J, Verhoog S, Bohlius J, Bramer W, et al. A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. Eur J Epidemiol. 2020;35(1):49-60. doi:10.1007/s10654-019-00576-5
14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. J Clin Epidemiol. 2021;134:103-12. doi:10.1016/j.jclinepi.2021.02.003
15. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. doi:10.1136/bmj.m689
16. Graafsma J, Murphy RM, van de Garde EMW, Karapinar-Çarkit F, Derijks HJ, Hoge RHL, et al. The use of artificial intelligence to optimize medication alerts generated by clinical decision support systems: A scoping review. J Am Med Inform Assoc. 2024;31(6):1411-22. doi:10.1093/jamia/ocae076
17. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: Systematic review. J Med Internet Res. 2025;27:e69678. doi:10.2196/69678
18. Jackson NJ, Brown KE, Miller R, Murrow M, Cauley MR, Collins BX, et al. Factors influencing the effectiveness of artificial intelligence-assisted decision-making in medicine: A scoping review. J Am Med Inform Assoc. 2026;33(5):1054-64. doi:10.1093/jamia/ocag002
19. Marcilly R, Colliaux J, Robert L, Pelayo S, Beuscart JB, Rousselière C, et al. Improving the usability and usefulness of computerized decision support systems for medication review by clinical pharmacists: A convergent, parallel evaluation. Res Social Adm Pharm. 2023;19(1):144-54. doi:10.1016/j.sapharm.2022.08.012
20. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969
21. Bolton WJ, Wilson R, Gilchrist M, Georgiou P, Holmes A, Rawson TM. The impact of artificial intelligence-driven decision support on uncertain antimicrobial prescribing: A randomised, multimethod study. Lancet Digit Health. 2025;7(11):100912. doi:10.1016/j.landig.2025.100912
22. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R; HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak. 2017;17(1):36. doi:10.1186/s12911-017-0430-8
23. Hussain MI, Reynolds TL, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: A systematic review. J Am Med Inform Assoc. 2019;26(10):1141-9. doi:10.1093/jamia/ocz095
24. Chen CY, Chen YL, Scholl J, Yang HC, Li YCJ. Ability of machine-learning based clinical decision support system to reduce alert fatigue, wrong-drug errors, and alert users about look alike, sound alike medication. Comput Methods Programs Biomed. 2024;243:107869. doi:10.1016/j.cmpb.2023.107869
25. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: Susceptibility in deployment of clinical decision-aids. NPJ Digit Med. 2021;4(1):31. doi:10.1038/s41746-021-00385-9
26. Kücking F, Busch DA, Przysucha M, Kutza JO, Hannemann N, Hüsers J, et al. Impact of AI recommendation correctness on diagnostic accuracy in clinical decision-making. Int J Med Inform. 2026;207:106223. doi:10.1016/j.ijmedinf.2025.106223
27. Rezaeian O, Asan O, Bayrak AE. The impact of AI explanations on clinicians' trust and diagnostic accuracy in breast cancer. Appl Ergon. 2025;129:104577. doi:10.1016/j.apergo.2025.104577
28. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: Clinicians' perspectives on trust, trustworthiness, and liability. Med Law Rev. 2023;31(4):501-20. doi:10.1093/medlaw/fwad013
29. Kashyap S, Morse KE, Patel B, Shah NH. A survey of extant organizational and computational setups for deploying predictive models in health systems. J Am Med Inform Assoc. 2021;28(11):2445-50. doi:10.1093/jamia/ocab154
30. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. NPJ Digit Med. 2020;3(1):17. doi:10.1038/s41746-020-0221-y
31. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
32. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-50. doi:10.1016/S2589-7500(21)00208-9
33. Goh E, Gallo RJ, Strong E, Weng Y, Kerman H, Freed JA, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. Nat Med. 2025;31(4):1233-8. doi:10.1038/s41591-024-03456-y
34. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33. doi:10.1038/s41591-022-01772-9
35. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nat Med. 2020;26(9):1364-74. doi:10.1038/s41591-020-1034-x
36. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nat Med. 2020;26(9):1351-63. doi:10.1038/s41591-020-1037-7