Review Article

Can Artificial Intelligence Distinguish Helpful Complexity from Harmful Polypharmacy? A Critical Evidence Review

Fatima El Idrissi[1]*, Samira Bennani[1], Youssef Benali[2], Ahmed Ben Ali[3]

[1]Department of AI and Polypharmacy Assessment, Faculty of Pharmacy, University of Marrakech, Marrakech, Morocco. [2]Department of Helpful vs. Harmful Complexity, Faculty of Pharmacy, University of Fez, Fez, Morocco. [3]Department of Clinical Pharmacy and AI Review, Faculty of Pharmacy, University of Casablanca, Casablanca, Morocco.

Abstract

Polypharmacy may represent necessary treatment for multimorbidity or expose patients to avoidable harm. Artificial intelligence is proposed to identify inappropriate medicines, predict interactions, and support deprescribing, but model accuracy does not establish whether a regimen's complexity is beneficial or harmful. To critically examine whether current systems distinguish helpful medication complexity from harmful polypharmacy, and to identify the evidence required for that claim. A bounded critical evidence review used seven frozen Google Scholar searches, supplementary citation searching, and publisher, PubMed, and Crossref verification. Of 104 captured records, 28 duplicates were removed, 76 were screened, 45 reports were sought, 43 were assessed, and 21 reports representing 21 studies or reviews were included. Evidence was extracted by construct, target, reference standard, validation design, patient context, clinical utility, and risk of inappropriate simplification. Prediction studies, interventions, reviews, and contextual studies were appraised using appropriate domains and synthesized without pooling. Most systems predicted criteria-defined potentially inappropriate medication labels, compiled rule-based warnings, estimated interaction or hospitalization risk, or generated clinician-mediated deprescribing suggestions. Prospective systems could increase discontinuation or reduce medication burden, yet clinical-outcome findings were limited and mixed. Patient goals, prognosis, indication, time-to-benefit, and treatment burden were rarely represented directly. Proxy-label circularity, restricted external validation, and outcome disconnect limited interpretation. Current evidence supports artificial intelligence as a source of partial medication-risk signals, not as a demonstrated judge of patient-specific regimen value. Distinguishing helpful complexity from harm requires explicit benefit representation, contextual goals, counterfactual comparison, prospective evaluation, human adjudication, and safety monitoring.

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

Introduction

Polypharmacy is often defined numerically, yet the same medicine count can describe fundamentally different clinical states. Medication count is an unstable proxy because definitions, prevalence estimates, and observed adverse associations vary with thresholds, time windows, morbidity, and study design [1-3]. A regimen containing many therapies may be proportionate when each medicine addresses a meaningful indication, provides a plausible benefit within the patient's expected time horizon, and remains acceptable in burden and risk. Conversely, a smaller regimen may still be harmful if it contains duplications, contraindicated combinations, poorly monitored high-risk medicines, or treatments that no longer serve the patient's priorities.

The central problem is therefore not whether artificial intelligence can count medicines or detect statistical associations. It is whether a system can identify the patient-specific balance between therapeutic necessity and avoidable burden while preserving uncertainty about diseases, preferences, prognosis, adherence, and competing outcomes. This judgment is also dynamic: a treatment may be appropriate during acute instability, become burdensome after recovery, or regain value when symptoms, life expectancy, care setting, or therapeutic goals change over time.

This distinction matters because contemporary systems operate through heterogeneous tasks that are frequently grouped under the same label. Some predict criteria-defined potentially inappropriate medication use; some identify drug–

Address for correspondence: Fatima El Idrissi, Department of AI and Polypharmacy Assessment, Faculty of Pharmacy, University of Marrakech, Marrakech, Morocco.
fatima.idrissi@uca.ma
Received: 14 April 2026; **Accepted:** 09 August 2026

How to cite this article: El Idrissi F, Bennani S, Benali Y, Ben Ali A. Can Artificial Intelligence Distinguish Helpful Complexity from Harmful Polypharmacy? A Critical Evidence Review. Arch Pharm Pract. 2026;17(3):36-45. https://doi.org/10.51847/SZ1ZvPQgAj

drug or drug–disease interactions; some estimate hospitalization or adverse-event risk; others assemble rule-based recommendations for medication review or deprescribing. Each task may contribute useful information, but none automatically answers whether the entire regimen is beneficial. A Beers or STOPP flag identifies a potentially inappropriate prescription under specified conditions, not proven net harm for every patient. An interaction model predicts a relation between drugs, not whether the combination's expected benefit outweighs that risk. A deprescribing prompt can change prescribing behavior without demonstrating improved symptoms, function, safety, or survival. These distinctions are especially important when simplification itself can cause harm through disease destabilization, withdrawal, loss of preventive benefit, or displacement of patient priorities.

This critical evidence review therefore asks six linked questions: how harmful polypharmacy is operationalized; which systems rely on medication count or circular proxy labels; whether models represent appropriate complexity, patient goals, prognosis, and treatment burden; what evidence supports deprescribing or clinical usefulness; which errors could produce inappropriate simplification; and what evidence is required before claiming that artificial intelligence distinguishes complexity from harm. The review separates retrospective model performance, prospective workflow effects, prescribing changes, and patient outcomes. Its interpretive standard is deliberately demanding: a clinically meaningful distinction requires more than classification against existing criteria. It requires evidence that the system represents benefits as well as risks, compares plausible alternatives, incorporates preference-sensitive context, communicates uncertainty, and supports monitored decisions. Deprescribing itself remains an individualized process requiring benefit–harm assessment, prognosis, goals, shared decision-making, and follow-up rather than automatic medication reduction [4].

Review Design, Questions, and Protocol

This review used a transparent critical evidence-review design to examine how constructs, labels, reference standards, validation strategies, and outcome choices shape claims about artificial intelligence and polypharmacy. The protocol was frozen after pilot searching and before final record capture. Six questions addressed operational definitions of harmful polypharmacy, reliance on medication-count or criterion-derived proxies, representation of benefit and patient context, evidence for deprescribing and clinical usefulness, risks of inappropriate simplification, and requirements for stronger claims. Record accounting distinguished records, reports, and represented studies [5].

Google Scholar was the primary discovery platform because the approved protocol specified a bounded Scholar review. Seven searches were run on 3 August 2026 using combinations of artificial intelligence, machine learning, polypharmacy, harmful polypharmacy, deprescribing recommendation systems, medication appropriateness, Beers or STOPP labels, drug interactions, patient goals, prognosis, and treatment burden. Backward and forward citation searching supplemented the captured records, while publisher, PubMed, and Crossref pages supported metadata and DOI verification. Reporting elements for broad evidence mapping were adapted without relabelling the study as a scoping review [6].

The review was designed to compare heterogeneous evidence rather than estimate a pooled intervention effect. Eligible evidence included prediction models, algorithmic or computerized medication-support systems, deprescribing interventions, contextual studies concerning multimorbidity or patient priorities, and reviews that directly informed interpretation. The synthesis focused on the relationship between what a system was trained or programmed to identify and what clinical claim was subsequently made. Critical review methodology was selected because explicit purpose, transparent selection, and structured interpretation can integrate diverse evidence while preserving methodological differences [7].

Review-level sources were evaluated for protocol clarity, search adequacy, selection transparency, appraisal, synthesis logic, and overinterpretation. These judgments informed evidential weight rather than producing a single numerical quality score. AMSTAR 2 concepts were used to identify weaknesses in secondary evidence without treating every included review as a formal intervention systematic review [8].

Eligibility, Definitions, and Information Sources

Reports published from 2017 through 2026 were eligible when they examined adults, especially older adults or people with multimorbidity, and supported a patient-level medication decision or safety interpretation. Systems had to classify, predict, detect, or recommend action concerning medication appropriateness, potentially inappropriate medicines, interactions, deprescribing, treatment burden, or polypharmacy consequences. Purely molecular interaction prediction without a medication-use context, medication count as the sole harm target, non-peer-reviewed work, and reports lacking interpretable targets or reference standards were excluded.

Medication count denoted the number of concurrent medicines and was not treated as harm. Regimen complexity included dosing, administration, monitoring, and organizational burden. A potentially inappropriate medication was a criterion- or judgment-defined treatment for which risk may outweigh benefit in specified circumstances. Beers criteria were treated as screening rules for potentially inappropriate use, not as direct measurements of patient-specific net harm [9]. Drug–drug and drug–disease interactions were extracted separately from whole-regimen appropriateness, while deprescribing opportunity meant a

reason to consider withdrawal rather than a mandatory discontinuation decision.

START omissions and STOPP-defined potentially inappropriate prescriptions were recorded as distinct constructs because underuse and overuse can coexist in the same regimen. Their application requires information about diagnosis, indication, duration, response, contraindications, and patient circumstances [10]. Helpful complexity was operationally defined for this review as treatment multiplicity that remains justified by expected benefit, goals, prognosis, tolerability, feasibility, and monitoring. Harmful polypharmacy required evidence of avoidable net burden, risk, or discordance with goals, rather than medicine count alone. Review design and eligibility boundaries were therefore driven by the questions and heterogeneity of the available evidence [11].

Search, Selection, Extraction, Appraisal, and Synthesis

Two reviewers screened captured titles and abstracts, sought full reports, assigned one primary full-text exclusion reason, and resolved disagreements through discussion. Extraction recorded study design, population, setting, system type, inputs, target or label, reference standard, comparator, validation context, key findings, clinical utility, representation of goals or prognosis, and principal applicability boundary. Prediction studies were appraised across participant selection, predictors, outcome definition, analysis, and applicability, using PROBAST-informed domains rather than an automated score [12].

Intervention studies were examined for allocation, deviations, outcome ascertainment, missing data, and selective interpretation; observational studies were assessed for confounding and transportability. Reviews were appraised using protocol, search, selection, and synthesis domains. Across designs, additional judgments addressed construct validity, criterion circularity, external validation, prospective utility, human mediation, patient-context representation, and whether a reported improvement concerned model performance, prescribing behavior, medication burden, or patient outcomes. Domain judgments and their supporting reasons were displayed rather than collapsed into an unsupported summary grade [13].

Evidence was grouped by decision function: field-level reviews, algorithmic support and PIM prediction, deprescribing interventions, contextual evidence, risk prediction, and technical interaction modeling. Contradictory findings were retained and interpreted through differences in targets, comparators, settings, follow-up, and outcome level. No pooled effect was calculated because interventions, labels, populations, and endpoints were too heterogeneous. Direction, consistency, evidential level, and limitations were synthesized narratively in accordance with synthesis-without-meta-analysis principles [14]. The bounded Scholar source, English-language interface, unstable platform ranking, and incomplete access to two reports limit comprehensiveness in practice; findings define evidence boundaries rather than exhaustive prevalence estimates.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for eligibility, definitions, and evidence-extraction framework for the review Can Artificial Intelligence Distinguish Helpful Complexity from Harmful Polypharmacy? A Critical Evidence Review.

Table 1. Eligibility, definitions, and evidence-extraction framework for the review Can Artificial Intelligence Distinguish Helpful Complexity from Harmful Polypharmacy? A Critical Evidence Review.

Review element	Operational definition	Eligibility or decision rule	Data field	Rationale	Bias or ambiguity risk	Planned synthesis use
Medication count	Number of concurrently documented medicines at a stated time point	Eligible only when connected to appropriateness, burden, interaction, safety, or another clinically interpretable construct	Medicine count; threshold; prescription status; observation window	Numerical definitions vary substantially and cannot independently establish harm [1].	Threshold dependence; incomplete medication lists; temporal instability	Separate descriptive exposure from harmful-polypharmacy claims
Regimen complexity	Review-defined combination of medication number, dosing, administration, monitoring, and organizational workload	Include when complexity is evaluated beyond medicine count or connected to patient feasibility	Dosing frequency; formulation; administration requirements; monitoring; adherence demands	Complexity may be justified when anticipated benefits remain proportionate to burden [4].	Complexity may be inferred from count rather than measured directly	Distinguish potentially helpful treatment intensity from avoidable burden
Potentially inappropriate medication	Medicine meeting a stated criterion or expert-judgment rule under specified circumstances	Include only when the criterion version, implementation, and target are identifiable	Criterion; trigger; exception; affected medicine; predicted or observed label	Beers criteria are screening criteria for potentially inappropriate use rather than direct net-harm measures [9].	Reference-standard circularity; missing exceptions; criterion transportability	Evaluate whether models reproduce proxy labels or identify patient-specific harm

Potential prescribing omission	Absence of an indicated medicine identified through an explicit omission criterion	Extract separately from medication excess and discontinuation opportunities	Omitted therapy; diagnosis; eligibility condition; contraindication	START criteria demonstrate that undertreatment may coexist with polypharmacy [10].	Incomplete diagnosis or contraindication data; presumption of universal benefit	Prevent simplification from being equated with better prescribing
Drug–drug or drug–disease interaction	Review-defined potential relationship between medicines, or between a medicine and a clinical condition, that may alter benefit or risk	Include when linked to patient-level medication assessment; exclude purely molecular prediction without clinical context	Interacting entities; mechanism; severity; evidence level; patient modifiers	Interaction detection supplies a risk signal but does not determine whole-regimen utility	Alert burden; database disagreement; omission of indication and competing benefit	Separate pairwise risk detection from regimen-level appropriateness
Deprescribing opportunity	A reason to consider dose reduction, substitution, or discontinuation, subject to clinician and patient review	Include when recommendation logic, mediation, and observed action are reported	Recommendation; rationale; accepting decision-maker; action; tapering; follow-up	Deprescribing requires individualized benefit–harm assessment and monitoring [4].	More discontinuation may be misinterpreted as improved care	Compare recommendation generation, prescribing change, and patient outcomes
Patient goals, prognosis, and treatment burden	Explicitly represented preferences, priorities, expected time horizon, and practical or experiential workload	Extract whether directly measured, clinician entered, inferred, or absent	Goal; preference; prognosis; time-to-benefit; burden; decisional role	These variables are central to determining whether complexity remains worthwhile [4].	Inaccurate inference; outdated preferences; exclusion of caregiver perspectives	Assess whether systems contain the context needed to identify helpful complexity
Clinical usefulness	Review-defined evidence that a system improves decisions, workflow, medication outcomes, safety, function, symptoms, utilization, or other patient-relevant outcomes	Model discrimination alone was insufficient; prospective or comparative evidence was extracted separately	Outcome level; comparator; follow-up; effect direction; uncertainty; adverse consequences	The review question requires separation of technical performance from decision and outcome performance	Surrogate substitution; short follow-up; confounding; selective outcome reporting	Construct a staged distinction between technical capability and demonstrated utility
Prediction-model appraisal	Domain-based assessment of participants, predictors, outcome, analysis, and applicability	Apply to every eligible prediction or warning model	Sampling; missing data; events per variable; validation; calibration; transportability	PROBAST-informed appraisal identifies bias and applicability concerns without reducing them to one score [12].	Incomplete reporting; post hoc judgments; inappropriate model comparisons	Interpret performance in relation to target validity and external applicability
Review and synthesis appraisal	Design-appropriate examination of search, selection, extraction, appraisal, grouping, and interpretation	Apply to included reviews and to the review's own non-pooled synthesis	Search scope; inclusion logic; appraisal method; synthesis grouping; certainty boundary	AMSTAR 2 concepts support review appraisal [8]; SWiM principles support transparent synthesis without pooling [14].	Heterogeneous review purposes; false equivalence across study designs	Preserve contradictions, explain evidential weight, and avoid unsupported pooled conclusions

Search and Selection Results

The frozen searches captured 95 Google Scholar records and nine supplementary citation-search records, producing 104 identified records. Deduplication removed 28 records, leaving 76 for title-and-abstract screening. Thirty-one records were excluded, 45 reports were sought, two could not be retrieved, and 43 full texts were assessed. Twenty-two full texts were excluded: five lacked an eligible artificial-intelligence, computerized decision-support, or analytically relevant intervention; five lacked patient-level medication-decision context; three treated medication count as the sole target; three were non-journal or preprint reports; two inadequately described targets or reference standards; two fell outside the year, quartile, or DOI boundary; and two were duplicate or non-contributing secondary reports. Twenty-one reports representing 21 distinct studies or reviews entered synthesis. Field reviews consistently describe a sparse and heterogeneous evidence base in which many systems remain rule-based or technically evaluated and person-centered clinical validation is limited [15-17].

Figure 1 presents the completed review-selection pathway in a black-and-white prisma-style flow diagram for evidence selection, documenting identification, deduplication, screening, eligibility, exclusions with reasons, and final inclusion using the values approved in the Part 1 PRISMA Record-Accounting Audit.

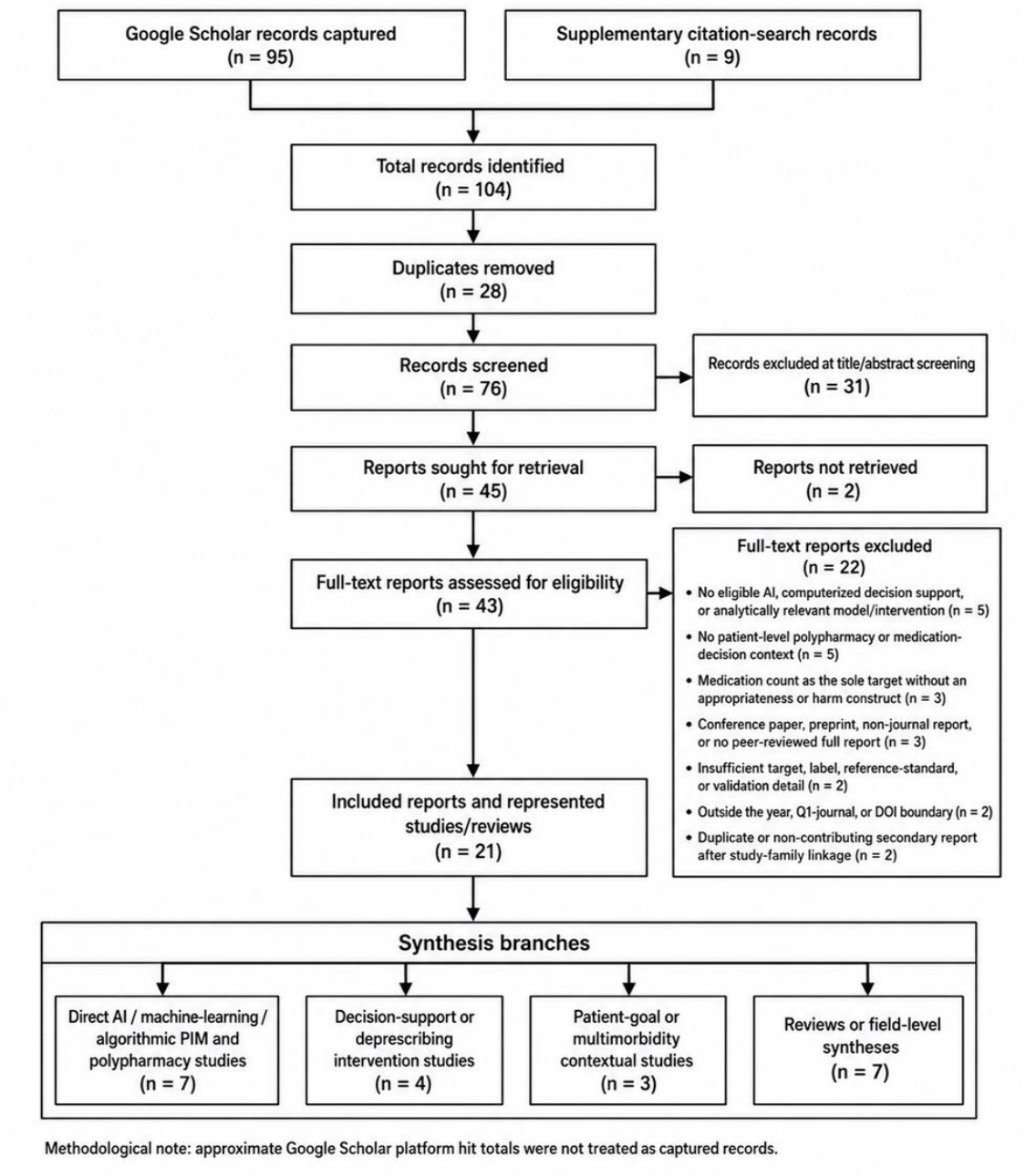

Figure 1. PRISMA-Style Flow Diagram for Evidence Selection

The included corpus divided arithmetically into seven direct algorithmic or machine-learning reports, four decision-support or deprescribing interventions, three patient-context studies, and seven reviews or field-level syntheses. These branches were descriptive rather than mutually ranked levels of evidence. Reviews mapped definitions and gaps; model studies tested prediction or alert generation; intervention studies examined changes in decisions or outcomes; contextual studies clarified variables needed to interpret medication value. No quantitative synthesis was undertaken because the branches used incompatible constructs and endpoints. Accordingly, the flow diagram documents selection and synthesis membership without implying that all included reports answered the central question with equal directness.

Harmful polypharmacy was not operationalized consistently across the included evidence. Some reports treated criterion-defined potentially inappropriate medication use as the principal target; others evaluated interaction warnings, deprescribing opportunities, medication burden, hospitalization risk, or broad adverse outcomes. This heterogeneity matters because each target answers a different question. A criterion flag indicates rule-defined concern, whereas hospitalization prediction estimates future risk and a deprescribing recommendation proposes an action requiring contextual review. The review therefore found no common

reference standard that directly measured the net patient-specific value of an entire regimen.

Proxy labels were especially prominent in prediction studies. When the outcome was generated from Beers, STOPP, or related rules, high discrimination primarily showed that a model could reproduce the chosen criteria from available variables. Such evaluation can support efficient screening, but it is partly circular when the model's clinical value is inferred from agreement with the same proxy used to create its labels. Missing indications, treatment response, previous intolerance, patient priorities, and time-to-benefit further restrict whether a positive label signifies avoidable harm.

Evidence-Base Characteristics

The included evidence comprised seven direct artificial-intelligence or machine-learning studies, four decision-support or deprescribing interventions, three patient-context studies, and seven reviews or field-level syntheses. Systems ranged from knowledge-based assistants and criteria-driven warnings to supervised prediction models and graph-based interaction algorithms. Most were developed or retrospectively tested; fewer were prospectively evaluated, and independent demonstrations of patient-centered clinical utility were uncommon.

Across system types, validation maturity formed a practical gradient. Development and internal testing established whether software functioned or reproduced labels. Retrospective comparative validation examined discrimination or classification within available data. Prospective workflow evaluation tested whether recommendations reached clinicians and changed actions. Patient-centered clinical evaluation asked whether those actions improved outcomes without creating new harms. Independent external validation and monitored implementation were rarely demonstrated. The review treated movement along this gradient as evidence of broader tested capability, not as a presumption that later stages were automatically superior or that an internally accurate model was clinically ready.

One web assistant assembled drug-interaction, potentially inappropriate medication, and rational-use information from embedded databases and rules, but its development evaluation did not establish patient-specific net benefit [18].

A related nursing-home and primary-care evaluation examined practical use of a rational-drug-use assistant. It suggested that algorithmically organized information could alter review processes and bring medication concerns to clinicians' attention. However, the observational context, limited control over confounding, and absence of decisive patient-centered endpoints prevented attribution of improved health outcomes to the system [19]. These reports illustrate a broader classification problem: an "AI-supported" label may describe rule compilation, information retrieval, or workflow presentation rather than learning, counterfactual reasoning, or autonomous medication judgment.

Prediction evidence focused more directly on classification. A warning model for older Chinese outpatients used clinical and medication variables to predict criterion-defined potentially inappropriate medicine labels. Its discrimination was relevant to case finding within the studied setting, yet the target remained a proxy produced by prescribing criteria rather than observed medication harm or preference-sensitive net benefit [20]. Consequently, performance could not establish that the model distinguished necessary complexity from inappropriate treatment.

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for characteristics and classification of the evidence included in the review Can Artificial Intelligence Distinguish Helpful Complexity from Harmful Polypharmacy? A Critical Evidence Review.

Table 2. Characteristics and classification of the evidence included in the review Can Artificial Intelligence Distinguish Helpful Complexity from Harmful Polypharmacy? A Critical Evidence Review.

Evidence category	Study or source design	Population or setting	AI system or intervention	Comparator or reference standard	Outcome or construct	Validation level	Key limitation
Field-level big-data evidence	Scoping review	Patient polypharmacy across heterogeneous settings	Big-data and analytical approaches	Study-defined targets and narrative comparison	Polypharmacy-management evidence landscape	Review-level evidence mapping	Few directly relevant patient-level studies and limited clinical evaluation [15]
Polypharmacy decision-support systems	Scoping review	Older adults with polypharmacy	Predominantly rule-based clinical decision-support systems	Rules, prescribing criteria, or expert judgment	Medication review, alerts, and recommendations	Technical and process evaluation predominated	Clinical outcomes and person-centered validation were uncommon [16]
AI-supported deprescribing landscape	Narrative review	Older patients receiving multiple medicines	Artificial-intelligence and algorithmic deprescribing tools	Mixed criteria and clinical approaches	Deprescribing support and prospective implementation potential	Review-level synthesis	Clinical integration, independent validation, and patient-centered evidence remained limited [17]

Web-based medication-risk assistant	System-development study	Geriatric medication users	Knowledge-based web assistant compiling drug-interaction, PIM, and rational-use information	Embedded databases and medication-safety rules	Warning and recommendation generation	Development and functional testing	No prospective clinical-outcome validation or direct net-benefit assessment [18]
Rational-drug-use assistant in practice	Observational system evaluation	Nursing-home patients and primary-care practice	AI-aided medication-review web assistant	Existing review processes and embedded knowledge rules	Medication assessment and workflow support	Single-context observational evaluation	Confounding and limited patient-centered outcome evidence restricted causal interpretation [19]
Criterion-defined PIM prediction	Machine-learning prediction study	Older Chinese outpatients in tertiary hospitals	Supervised warning model using clinical and medication variables	Criteria-derived PIM labels	Prediction of potentially inappropriate medicine use	Internal validation in a restricted setting	Label circularity, limited transportability, and no prospective decision or outcome evaluation [20]

Deprescribing Recommendation Systems

Prospective interventions provided stronger evidence about decision effects than retrospective prediction studies. Electronic support could identify deprescribing opportunities, communicate recommendations, and increase discontinuation, while multidisciplinary transition interventions could reduce medication burden. Nevertheless, recommendations were mediated by clinicians and, where relevant, patients or caregivers; systems did not autonomously determine net regimen value. Prospective interventions increased deprescribing actions or reduced medication burden, although clinical-outcome evidence remained mixed and the decisions were clinician mediated [21-23].

The controlled MedSafer study demonstrated that individualized electronic decision support could increase recognition and action on deprescribing opportunities in hospitalized older adults [21]. The subsequent multicenter cluster-randomized trial more clearly separated process effects from patient outcomes: potentially inappropriate medicine discontinuation increased, but 30-day adverse drug events were not significantly reduced [22]. This result does not negate value in medication review, but it prevents prescribing change from being treated as proof of improved safety.

Shed-MEDS used a human-intensive, multidisciplinary process extending from hospitalization into postacute care. Medication burden fell, and patient or caregiver priorities informed decisions, but the intervention was not autonomous artificial intelligence [23]. Together, these studies show that systems can facilitate action while leaving the central judgment distributed among software, pharmacists, physicians, patients, and follow-up processes.

Patient-Goal and Prognosis Representation

Recognizing helpful complexity requires representing why each medicine exists, what outcome it serves, when benefit is expected, and whether that outcome matters to the patient. Patient Priorities Care showed that goal-aligned decision-making could reduce treatment burden and healthcare tasks while maintaining attention to what patients considered important [24]. This study did not validate an artificial-intelligence system; it identified contextual variables that such a system would need to handle.

Deprescribing decisions also varied across clinical cases and countries in a 31-country vignette study, indicating that judgment depends on comorbidity, symptoms, medicine purpose, risk tolerance, and professional interpretation rather than on medication count alone [25]. Variation may reflect inconsistent practice, but it can also represent legitimate preference-sensitive tradeoffs that a fixed proxy label cannot resolve.

Observed multimorbidity and polypharmacy patterns further showed that medicine combinations co-vary with disease patterns [26]. A complex regimen may therefore encode appropriate responses to multiple conditions as well as accumulated prescribing burden. Most reviewed systems represented diagnoses and medicine lists more readily than goals, prognosis, caregiver capacity, adherence burden, or changing preferences. Evidence for clinical usefulness was strongest for supporting review and prescribing action, weaker for sustained patient outcomes, and insufficient for autonomous distinction between helpful and harmful complexity.

Risks of Inappropriate Simplification

Inappropriate simplification can arise when a system treats every proxy-positive medicine as removable, overlooks under-treatment, ignores indication or time-to-benefit, or translates population associations into individual causal decisions. Technical sophistication does not eliminate these risks. Potential harms extend beyond excessive discontinuation. A false-positive recommendation may remove symptomatic or preventive benefit; a false negative may preserve avoidable risk; and poorly timed withdrawal may cause rebound, destabilization, or loss of treatments

aligned with patient priorities. In one comparative study, machine-learning approaches did not clearly outperform traditional regression for predicting potentially inappropriate medicine use [27]. Added complexity therefore did not necessarily produce added decision value.

A dementia-specific model identified criterion-defined potentially inappropriate medicine use within a focused population, but its interpretation remained bounded by the label, setting, and available predictors [28]. Transporting such a model without recalibration or contextual exceptions could amplify misclassification in populations with different disease severity, goals, prescribing cultures, or data completeness.

Clinical interventions also caution against assuming that structured optimization automatically improves hard outcomes. OPERAM integrated medication review and contextual clinical information, yet did not significantly reduce drug-related hospital admissions [29]. The finding supports uncertainty rather than intervention failure: medication outcomes are influenced by baseline risk, implementation fidelity, competing causes, and follow-up.

The review-derived evidence requirements are therefore explicit target validity, representation of indication and expected benefit, patient goals and prognosis, treatment burden, alternatives, uncertainty, and monitoring. Evaluation should compare plausible medication plans rather than only predict a proxy label, and should measure accepted recommendations, inappropriate discontinuations, under-treatment, symptoms, function, safety, utilization, and equity. These requirements are an original critical synthesis, not a validated clinical rule or deployment specification.

Discussion: Principal Findings and Interpretation

The central answer is qualified but clear. Current artificial-intelligence and algorithmic systems can detect selected medication-risk signals, reproduce prescribing criteria, organize warnings, and support clinician-mediated deprescribing. They have not demonstrated a general capacity to determine whether an individual patient's entire regimen represents helpful complexity or harmful polypharmacy. That distinction requires causal and value-sensitive reasoning about each medicine's indication, expected benefit, competing risks, time horizon, burden, alternatives, interactions, and relationship to patient priorities. A hospitalization-risk model can identify patients needing review, but even when potentially inappropriate prescribing features contribute to prediction, it does not identify which medication should be changed or whether discontinuation would improve the outcome [30].

The evidence therefore supports a modular interpretation of capability. Interaction and adverse-effect models may contribute a hazard module; PIM classifiers may contribute a criteria-screening module; prognosis and treatment-burden assessments may contribute contextual modules; and deprescribing systems may organize candidate actions. Their outputs remain partial unless integrated around a patient-specific decision question. Explainable drug–drug interaction research can clarify why a pair is predicted to interact, yet the technical literature remains weakly connected to indications, disease control, competing benefit, patient preferences, and prospective workflow utility [31]. Explanation of a prediction is not equivalent to justification of a clinical action.

Graph models similarly extend the scale of detectable relationships without resolving regimen value. Polypharmacy side-effect prediction can infer associations among drug combinations and adverse effects, but it does not establish whether a given combination should be maintained, modified, monitored, or stopped for a particular person [32]. The missing step is counterfactual: what is expected to happen under the current regimen compared with feasible alternatives, including no change, dose adjustment, substitution, staged withdrawal, or addition of omitted therapy? Evidence about both benefits and harms must accompany these alternatives.

Five conditions emerge for a stronger claim. First, the target must represent patient-specific net value rather than medicine count or criterion agreement alone. Second, input data must include indication, treatment response, duration, adherence, renal and hepatic function, prognosis, time-to-benefit, prior withdrawal attempts, and patient or caregiver priorities. Third, the system must distinguish screening from recommendation and recommendation from final decision. Fourth, evaluation must proceed from internal validity to external validation, prospective comparative testing, and monitored implementation. Fifth, safety assessment must include inappropriate discontinuation, omission, disease destabilization, withdrawal, burden transfer, alert fatigue, and unequal performance across populations.

These conditions have direct implications for clinical pharmacy. Pharmacists are positioned to reconcile medicines with diagnoses, identify missing contextual data, challenge proxy-positive recommendations, and translate model outputs into monitored options. Artificial intelligence may prioritize records, surface interacting factors, or produce structured questions, while pharmacists determine whether the evidence applies and whether the patient accepts the proposed tradeoff. Systems should expose the provenance of criteria, the variables driving predictions, missing information, confidence, applicable population, and circumstances requiring escalation.

Evaluation should also compare the system with realistic alternatives. A model may outperform no review while adding little beyond a pharmacist using established criteria, or it may improve consistency without improving outcomes. Comparators should include usual care, structured pharmacist review, conventional statistical models, and simpler rule-based systems. Outcomes should be organized hierarchically: technical performance; recommendation quality; clinician

acceptance; medication changes; patient-reported burden; symptoms and function; adverse drug events; hospital use; mortality; and unintended consequences. Longer follow-up is necessary when preventive benefits or withdrawal effects unfold slowly.

Implications, Strengths, and Limitations

The research agenda should begin with better labels. Observational associations between polypharmacy and mortality, falls, hospitalization, or functional decline are important for risk awareness, but they remain heterogeneous and vulnerable to confounding by multimorbidity, frailty, indication, and healthcare contact [33]. Training a model on such outcomes without a causal decision framework may identify vulnerability rather than modifiable medication harm. Future datasets should link medicine-level indications, reasons for continuation or withdrawal, patient priorities, adjudicated appropriateness, longitudinal outcomes, and consequences of alternative decisions.

Prospective studies should test whether systems improve the quality of medication decisions, not only whether they reduce counts or generate accepted alerts. Existing synthesis suggests that deprescribing interventions more consistently reduce medication burden than improve clinical outcomes, with effects varying across populations, interventions, and endpoints [34]. Trials should prespecify benefits and harms, document clinician and patient mediation, measure under-treatment and inappropriate discontinuation, and analyze subgroup performance. Pragmatic implementation studies should evaluate data quality, alert burden, workflow displacement, accountability, and post-deployment drift.

Governance should preserve a reversible, reviewable pathway from signal to action. High-risk recommendations should require documented indication review, patient-goal confirmation, tapering or substitution plans, and follow-up. Safe deprescribing remains an individualized process involving prioritization, shared decisions, withdrawal planning, and monitoring [35]. System outputs should therefore be treated as contestable evidence, with clear responsibility for adjudication and mechanisms for patients and clinicians to reject, modify, or revisit recommendations.

Strengths include explicit separation of medication count, complexity, PIMs, interactions, deprescribing opportunities, goals, prognosis, burden, and clinical usefulness; its distinction among model, decision, workflow, and outcome performance; and its preservation of contradictory findings. The record-level accounting also prevents platform hit estimates from being mistaken for screened records. Its limitations are substantial. Google Scholar was the sole primary discovery platform, ranking is unstable, English-language searching may omit relevant evidence, two reports were not retrieved, and heterogeneity precluded pooled estimates. Some included reports were contextual rather than direct artificial-intelligence evaluations, because they defined information required to judge helpful complexity. Appraisal relied on reported information and design-appropriate domains rather than a common validated score.

Accordingly, this review does not establish that artificial intelligence cannot eventually distinguish helpful complexity from harmful polypharmacy. It establishes that the reviewed evidence has not yet demonstrated that broad capability. The proposed requirements are an evidence-bounded research framework, not a clinical directive, regulatory standard, or deployment-ready architecture.

Conclusion

Current artificial-intelligence systems do not yet demonstrate a general ability to distinguish helpful medication complexity from harmful polypharmacy. The reviewed evidence shows narrower capabilities: reproducing potentially inappropriate medication criteria, identifying interaction or adverse-effect signals, estimating hospitalization risk, organizing medication information, and presenting clinician-mediated deprescribing opportunities. These functions may improve case finding or structure medication review, but they do not independently determine the patient-specific net value of a regimen.

The central limitation is conceptual as well as technical. Medication count, criterion-defined potentially inappropriate use, predicted interaction, treatment burden, and avoidable harm are related but non-equivalent constructs. A clinically meaningful judgment must consider why each medicine is used, its expected benefit and time-to-benefit, competing risks, disease control, prognosis, practical burden, patient goals, available alternatives, and consequences of changing treatment. Current studies rarely represent all these elements or evaluate them prospectively against patient-centered outcomes.

A stronger evidence base requires targets that reflect net medication value, transparent and transportable models, comparison with pharmacists and simpler decision-support methods, prospective evaluation, and measurement of both benefit and unintended harm. Recommendations should remain contestable, reversible, and subject to patient and professional adjudication, tapering plans, and follow-up monitoring. Independent validation across settings remains especially important. Artificial intelligence is therefore best understood at present as a source of partial evidence within medication review rather than an autonomous judge of regimen appropriateness. The proposed requirements define a research and evaluation agenda; they do not establish a validated clinical framework, universal decision rule, regulatory standard, or deployment-ready system.

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

References

1. Masnoon N, Shakib S, Kalisch-Ellett L, Caughey GE. What is polypharmacy? A systematic review of definitions. BMC Geriatr. 2017;17(1):230. doi:10.1186/s12877-017-0621-2
2. Morin L, Johnell K, Laroche ML, Fastbom J, Wastesson JW. The epidemiology of polypharmacy in older adults: Register-based prospective cohort study. Clin Epidemiol. 2018;10:289-98. doi:10.2147/CLEP.S153458
3. Wastesson JW, Morin L, Tan ECK, Johnell K. An update on the clinical consequences of polypharmacy in older adults: A narrative review. Expert Opin Drug Saf. 2018;17(12):1185-96. doi:10.1080/14740338.2018.1546841
4. Thompson W, McDonald EG. Polypharmacy and deprescribing in older adults. Annu Rev Med. 2024;75:113-27. doi:10.1146/annurev-med-070822-101947
5. 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. doi:10.1136/bmj.n71
6. 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
7. Snyder H. Literature review as a research methodology: An overview and guidelines. J Bus Res. 2019;104:333-9. doi:10.1016/j.jbusres.2019.07.039
8. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358. doi:10.1136/bmj.j4008
9. 2023 American Geriatrics Society Beers Criteria® Update Expert Panel. American Geriatrics Society 2023 updated AGS Beers Criteria® for potentially inappropriate medication use in older adults. J Am Geriatr Soc. 2023;71(7):2052-81. doi:10.1111/jgs.18372
10. O'Mahony D, Cherubini A, Renom Guiteras A, Denkinger M, Beuscart JB, Onder G, et al. STOPP/START criteria for potentially inappropriate prescribing in older people: Version 3. Eur Geriatr Med. 2023;14(4):625-32. doi:10.1007/s41999-023-00777-y
11. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143. doi:10.1186/s12874-018-0611-x
12. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-8. doi:10.7326/M18-1376
13. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods. 2021;12(1):55-61. doi:10.1002/jrsm.1411
14. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ. 2020;368. doi:10.1136/bmj.l6890
15. Wilfling D, Hinz A, Steinhäuser J. Big data analysis techniques to address polypharmacy in patients—a scoping review. BMC Fam Pract. 2020;21(1):180. doi:10.1186/s12875-020-01247-1
16. Mouazer A, Tsopra R, Sedki K, Letord C, Lamy JB. Decision-support systems for managing polypharmacy in the elderly: A scoping review. J Biomed Inform. 2022;130:104074. doi:10.1016/j.jbi.2022.104074
17. Al Meslamani AZ. Management of polypharmacy through deprescribing in older patients: A review of the role of AI tools. Expert Rev Clin Pharmacol. 2025;18(6):333-45. doi:10.1080/17512433.2025.2519648
18. Akyon SH, Akyon FC, Yılmaz TE. Artificial intelligence-supported web application design and development for reducing polypharmacy side effects and supporting rational drug use in geriatric patients. Front Med (Lausanne). 2023;10:1029198. doi:10.3389/fmed.2023.1029198
19. Yılmaz T, Ceyhan Ş, Akyön ŞH, Yılmaz TE. Enhancing primary care for nursing home patients with an artificial intelligence-aided rational drug use web assistant. J Clin Med. 2023;12(20):6549. doi:10.3390/jcm12206549
20. Hu Q, Tian F, Jin Z, Lin G, Teng F, Xu T. Developing a warning model of potentially inappropriate medications in older Chinese outpatients in tertiary hospitals: A machine-learning study. J Clin Med. 2023;12(7):2619. doi:10.3390/jcm12072619
21. McDonald EG, Wu PE, Rashidi B, Forster AJ, Huang A, Pilote L, et al. The MedSafer Study: A controlled trial of an electronic decision support tool for deprescribing in acute care. J Am Geriatr Soc. 2019;67(9):1843-50. doi:10.1111/jgs.16040
22. McDonald EG, Wu PE, Rashidi B, Goodwin Wilson M, Bortolussi-Courval E, Atique A, et al. The MedSafer Study—electronic decision support for deprescribing in hospitalized older adults: A cluster randomized clinical trial. JAMA Intern Med. 2022;182(3):265-73. doi:10.1001/jamainternmed.2021.7429
23. Vasilevskis EE, Shah AS, Hollingsworth EK, Shotwell MS, Kripalani S, Mixon AS, et al. Deprescribing medications among older adults from end of hospitalization through postacute care: A Shed-MEDS randomized clinical trial. JAMA Intern Med. 2023;183(3):223-31. doi:10.1001/jamainternmed.2022.6545
24. Tinetti ME, Naik AD, Dindo L, Costello DM, Esterson J, Geda M, et al. Association of patient priorities-aligned decision-making with patient outcomes and ambulatory health care burden among older adults with multiple chronic conditions: A nonrandomized clinical trial. JAMA Intern Med. 2019;179(12):1688-97. doi:10.1001/jamainternmed.2019.4235
25. Jungo KT, Mantelli S, Rozsnyai Z, Missiou A, Gerasimovska Kitanovska B, Weltermann B, et al. General practitioners' deprescribing decisions in older adults with polypharmacy: A case vignette study in 31 countries. BMC Geriatr. 2021;21(1):19. doi:10.1186/s12877-020-01953-6
26. Stafford G, Villén N, Roso-Llorach A, Troncoso-Mariño A, Monteagudo M, Violán C. Combined multimorbidity and polypharmacy patterns in the elderly: A cross-sectional study in primary health care. Int J Environ Res Public Health. 2021;18(17):9216. doi:10.3390/ijerph18179216
27. Chiu YM, Sirois C, Simard M, Gagnon ME, Talbot D. Traditional methods hold their ground against machine learning in predicting potentially inappropriate medication use in older adults. Value Health. 2024;27(10):1393-9. doi:10.1016/j.jval.2024.06.005
28. Hu Q, Zhao M, Teng F, Lin G, Jin Z, Xu T. A model for identifying potentially inappropriate medication used in older people with dementia: A machine learning study. Int J Clin Pharm. 2024;46(4):937-46. doi:10.1007/s11096-024-01730-0
29. Blum MR, Sallevelt BTGM, Spinewine A, O'Mahony D, Moutzouri E, Feller M, et al. Optimizing Therapy to Prevent Avoidable Hospital Admissions in Multimorbid Older Adults (OPERAM): Cluster randomised controlled trial. BMJ. 2021;374. doi:10.1136/bmj.n1585
30. Olender RT, Roy S, Nishtala PS. Potentially inappropriate polypharmacy is an important predictor of 30-day emergency hospitalisation in older adults: A machine learning feature validation study. Age Ageing. 2025;54(6). doi:10.1093/ageing/afaf156
31. Vo TH, Nguyen NTK, Kha QH, Le NQK. On the road to explainable AI in drug-drug interactions prediction: A systematic review. Comput Struct Biotechnol J. 2022;20:2112-23. doi:10.1016/j.csbj.2022.04.021
32. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13). doi:10.1093/bioinformatics/bty294
33. Davies LE, Spiers G, Kingston A, Todd A, Adamson J, Hanratty B. Adverse outcomes of polypharmacy in older people: Systematic review of reviews. J Am Med Dir Assoc. 2020;21(2):181-7. doi:10.1016/j.jamda.2019.10.022
34. Zhou D, Chen Z, Tian F. Deprescribing interventions for older patients: A systematic review and meta-analysis. J Am Med Dir Assoc. 2023;24(11):1718-25. doi:10.1016/j.jamda.2023.07.016
35. Hung A, Kim YH, Pavon JM. Deprescribing in older adults with polypharmacy. BMJ. 2024;385. doi:10.1136/bmj-2023-074892