

Research Article

Medication safety and artificial intelligence: evidence, failure modes, and how to vet AI tools before they touch patient care

Patricia Rizal^{1,*}¹ Davao Medical School Foundation Inc, Davao City, Philippines* Correspondence: patricia_rizal45@yahoo.com<https://doi.org/10.59652/hrnhkr35>

Abstract: Medication harm remains common, with a substantial portion preventable at the stages of prescribing and administration. (1) I reviewed evidence on how artificial intelligence (AI)—including rule-based clinical decision support, machine learning (ML), natural language processing (NLP), and large language models (LLMs)—can improve medication safety, and how these tools can fail. Methods: On 10 February 2026, I searched PubMed for systematic reviews/meta-analyses, high-quality observational studies, human factors evidence, reporting/risk-of-bias guidance, and peer-reviewed LLM evaluations relevant to prescribing error detection, medication reconciliation, dosing support, adverse drug event (ADE) prediction, pharmacovigilance, and medication-related documentation. All included citations were verified on PubMed by confirming PMIDs and bibliographic details. Results: Evidence for CPOE/CDSS suggests improvements in practitioner performance and some reductions in preventable ADEs, but effects are heterogeneous and systems can facilitate new error types. (2–5) Safety alerts are frequently overridden (49%–96%), and evidence that interruptive prescribing alerts improve patient outcomes is limited. (6,7) Automation bias is a documented risk: correct CDSS can reduce omission errors, yet incorrect CDSS can increase omission errors and induce commission errors. (12,13,15) ML-based ADE prediction shows emerging performance (pooled AUC 0.72, 95% CI 0.68–0.75), but prediction studies frequently show high risk of bias, motivating use of TRIPOD/PROBAST and careful validation. (9–11,17) In pharmacovigilance using routinely collected data, methods and performance reporting are heterogeneous, and no method is uniformly superior. (18) In medication text workflows, LLMs can reduce specific direction errors when designed with domain logic and guardrails, but hallucinations and omissions remain safety-critical. (20,21) Conclusions: AI can improve medication safety when tasks are bounded, outputs are auditable, and systems are integrated with strong human factors design and lifecycle governance; deployment requires local validation, bias and dataset shift monitoring, and formal safety evaluation—especially for generative AI. (9–11,20–27)

Keywords: Medication safety, AI tools, Medication harm

Received: 09 June 2025

Accepted: 12 December 2025

Published: 28 December 2025

**Copyright:** © 2022 by the authors.

Submitted for open access publication

under the terms and conditions of the

Creative Commons Attribution (CC BY)

license

(https://creativecommons.org/licenses/by/4.0/).

Introduction

Medication harm remains a high-frequency, high-impact patient safety problem, and a substantial fraction of adverse drug events (ADEs) is preventable at the stage of medication ordering and administration. (1) In the last two decades, electronic prescribing, computerised provider order entry (CPOE), and clinical decision support systems (CDSS) have been widely implemented to reduce preventable medication errors and preventable ADEs (pADEs), but the evidence shows heterogeneous effects and non-trivial unintended consequences. (2,3)

A central practical challenge is that “medication safety technology” is not a single intervention: it is a sociotechnical system that depends on alert content, workflow fit, user training, interface design, local data quality, and governance. Even when CPOE reduces certain classes of errors, problematic ordering pathways, default selections, fragmented displays, and brittle configuration can facilitate new errors. (4,5) This matters because medication safety is both a

technical and behavioural domain: clinicians and pharmacists must trust—yet actively verify—automated outputs under time pressure.

Safety alerts illustrate the trade-off. In a classic systematic review of physician response to drug safety alerts, override rates were high (typically reported between 49% and 96%), and error-producing conditions included low specificity, unclear content, and workflow disruption. (6) A later systematic review of interruptive medication prescribing alerts found that just over half of studied alerts produced statistically significant beneficial effects on prescriber behaviour, while patient outcomes were evaluated infrequently and did not clearly improve in the small subset of studies that measured them. (7) In other words, digitisation and decision support can improve the medication-use process, but they can also shift risk into new failure modes.

Artificial intelligence (AI) now enters this landscape in multiple forms: traditional rule-based CDSS (often not considered “AI” in current terminology), machine-learning (ML) models that predict risk or detect anomalies, natural language processing (NLP) systems that extract medication signals from text, and generative AI/large language models (LLMs) that draft, interpret, or verify medication-related text. Each brings distinct promises and distinct hazards. A medication-focused systematic review found evidence that medication-related CDSS can improve process and (in some cases) clinical outcomes, while also highlighting implementation factors that determine effectiveness and safety. (8)

From a clinician’s perspective, the question is not “Is AI good or bad?”. The operational question is: **Which AI applications for medication safety have credible evidence, what are the credible failure modes, and what minimum vetting is required before deployment?** This manuscript synthesises high-quality evidence across the medication-use process (prescribing, reconciliation, dosing, ADE prediction, pharmacovigilance, and patient-facing/documentation pathways) and proposes a pragmatic pre-deployment and post-deployment clinical vetting framework grounded in empirical evidence and reporting standards.

Methods

This is a narrative evidence synthesis with a medication-safety lens, targeting a manuscript length of ~3000–3500 words (default target proceeded with, as requested). I searched PubMed on **10 February 2026** and verified each included citation by confirming the PMID and bibliographic details on the PubMed record (PMID in the reference list; DOI included where available). Search terms combined medication-safety concepts with AI/CDSS concepts, for example: *“medication errors” AND (clinical decision support OR CPOE); “automation bias” AND prescribing; “dataset shift” AND clinical AI; “prediction model” AND risk of bias AND machine learning; “CONSORT-AI” OR “SPIRIT-AI”; “natural language processing” AND medication safety; “adverse drug event prediction” AND machine learning AND meta-analysis; “signal detection” AND routinely collected data AND systematic review; and “large language model” AND medication.*

This study prioritised (i) systematic reviews/meta-analyses and high-quality overviews for CDSS and ML prediction, (ii) empirically grounded human factors evidence for automation bias and alerting behaviour, (iii) reporting and risk-of-bias guidance for AI intervention trials and prediction models, and (iv) peer-reviewed evaluations of LLMs in medication-adjacent tasks. Where quantitative estimates were used, I quoted only values explicitly reported in the source (and cited them immediately). Where effect sizes were not consistently available or not comparable, I reported findings qualitatively.

Results

Evidence map and deployment reality

Across medication safety tasks, the evidence base clusters into three partially overlapping domains:

1. **Rule-based CDSS/CPOE evidence** (often pre-dating current “AI” terminology): relatively mature, demonstrates improved process measures and

sometimes reduced errors, but with wide heterogeneity and known unintended consequences. (2–7,8)

2. **ML prediction and detection models** (ADE prediction, risk stratification, anomaly detection): expanding quickly, but threatened by methodological fragility (overfitting, leakage, missing-data mishandling), poor reporting, and limited external/temporal validation. (9–13)
3. **NLP and LLM systems** (text extraction, reconciliation support, direction checking, drafting): promising for scale and speed, but vulnerable to hallucinations, silent omissions, poor grounding to authoritative sources, and workflow-induced automation bias. (14–18)

The key practical pattern is that **evidence strength is highest when the task is narrow, well-defined, and evaluable against a clear reference standard**, and weakest when the system's output is open-ended, difficult to verify quickly, or depends on local context (formularies, lab units, prescribing norms, handwriting/scans, multilingual patient instructions).

Prescribing error detection and medication-order decision support

Evidence signal. Systematic review evidence indicates that computerised clinical decision support can improve practitioner performance and, in some settings, patient outcomes, though the magnitude and reliability vary across interventions and contexts. (3) In hospital settings, a systematic review and meta-analysis evaluating CPOE found evidence that CPOE can reduce preventable ADEs, but effects on medication errors were heterogeneous and context-dependent. (2)

Failure modes. Two complementary empirical streams define medication-safety failure modes in CPOE/CDSS:

- **Technology-facilitated errors.** A landmark investigation of a leading CPOE system documented how system design could facilitate medication error risks, emphasising that implementers must attend to errors that systems cause in addition to those they prevent. (4)
- **Vulnerability and recurrence.** An analysis of medication error reports where CPOE was a “contributing cause” showed that error reports can identify recurring, testable scenarios to which contemporary CPOE systems remain vulnerable, supporting the need for enhanced monitoring, reporting, and vulnerability testing. (5)

Alerting behaviour and alert fatigue as safety mechanisms. Clinician response mediates CDSS effectiveness. High override rates are not necessarily “bad”—many alerts are non-actionable or irrelevant—but they can signal poor specificity, workflow disruption, or low trust. (6) In the systematic review of interruptive prescribing alerts, 53% of studies reported beneficial effects on prescriber behaviour, 34% reported no significant effect, and a small proportion reported detrimental effects; evidence for improved patient outcomes was limited, with only two studies evaluating patient outcomes and neither showing significant improvement. (7)

Interpretation for clinicians and pharmacists. For prescribing error detection and order support, I interpret the mature evidence base as supporting cautious optimism: CDSS can reduce certain errors, but safety depends on local configuration quality, continuous monitoring, and a realistic understanding that decision support **shifts** error patterns rather than eliminating error.

Medication reconciliation and medication history support

Medication reconciliation (MedRec) is a safety-critical process at transitions of care, where discrepancies can propagate into downstream prescribing and administration.

Electronic tools have been proposed to improve completeness and reduce discrepancies. In a systematic review of electronic tools to support medication reconciliation, included evaluations spanned quality improvement projects, observational studies, and a small number of randomised trials or trial protocols; evaluations largely focused on usability and implementation outcomes, and only one randomised trial evaluated effects on potential adverse drug events. (14)

Interpretation. The evidence suggests that eMedRec tools can improve process reliability, but robust evidence for patient-level harm reduction is still limited and heavily dependent on implementation context, data integration quality, and pharmacological expertise embedded in workflows. (14) From a vetting standpoint, reconciliation tools should be assessed not only for accuracy of the compiled medication list, but for (i) provenance transparency (what sources were used), (ii) conflict resolution logic, and (iii) usability under real admission/discharge time constraints.

Dosing decision support, calculation, and high-risk medication workflows

Dosing is one of the highest-stakes domains for decision support because errors can quickly translate into immediate harm, particularly in paediatrics and in renal/hepatic impairment. Classical CDSS dosing alerts can reduce omission errors when correct, but can also induce over-reliance. A controlled experimental study in e-prescribing demonstrated that correct CDSS reduced omission errors (e.g., reductions reported as 38.3% in low-complexity scenarios), while incorrect CDSS increased omission errors and produced substantial commission errors (acceptance of false positive alerts). (15) This illustrates a general medication-safety principle: **the cost of a “confidently wrong” recommendation can exceed the benefit of a “usually right” system when verification is difficult.**

Recent generative AI studies suggest that LLMs can perform well on constrained dosing calculations, but the scope of safe generalisation remains uncertain. In a cross-sectional evaluation comparing LLMs with experienced nurses on paediatric medication calculations, nurses averaged 93.14 (± 9.39) accuracy, while two LLMs achieved 100 accuracy; LLM response times were markedly faster than nurses. (16) These results are intriguing but should not be overgeneralised to broader dosing decision support, which often involves contextual clinical judgement (indication, comorbidities, interacting drugs, lab trends, formulation constraints, and monitoring). (16)

Interpretation. For dosing tools—especially those involving LLMs—I treat “calculator-like” tasks (unit conversion, weight-based dose calculation given a fixed rule) as potentially safer entry points, provided the system is bounded, auditable, and integrated with strong checking (e.g., displaying formulae and intermediate steps, forcing unit confirmation, and triggering mandatory human verification for high-alert meds). The same evidence also reinforces the need to formally evaluate automation bias and error recovery in realistic workflows. (15)

ADE prediction and risk stratification using machine learning

Machine learning for ADE prediction aims to identify patients at higher risk of harm during hospitalisation or therapy, potentially enabling preventive actions (review, monitoring, pharmacist consults, deprescribing, renal-dose adjustment). A systematic review and meta-analysis of ML models predicting multiple ADEs using EHR data reported a pooled estimated AUC of 0.72 (95% CI 0.68–0.75), based on 10 included studies. (17)

This level of discrimination is clinically meaningful only if (i) calibration is adequate, (ii) predictions are actionable, and (iii) the intervention triggered by prediction reduces harm without creating alert fatigue or inequitable resource allocation. Methodologically, ADE prediction models are susceptible to known pitfalls: outcome label bias (e.g., ADEs detected more often in patients who are monitored more intensively), confounding by indication, data leakage, and overfitting. These risks motivate formal reporting and bias assessment.

In prediction model research, reporting standards (TRIPOD) and risk-of-bias assessment (PROBAST) provide structured expectations for transparency, including details of model development, validation, calibration, and performance reporting. (9,10) A systematic

review using PROBAST across supervised ML prediction models found that a high proportion of analyses were at high risk of bias (reported as 87% of analysed models/validations), with frequent problems including inadequate events-per-predictor, poor missing data handling, and inadequate assessment of overfitting. (11)

Interpretation. For ADE prediction tools, I consider internal validation alone inadequate for deployment decisions. At minimum, local external validation (temporal and site-level), calibration checks, subgroup performance audits, and a clear “decision-curve” framing of clinical utility are required before integrating a model into medication workflows. (9–11)

Pharmacovigilance and signal detection

At population level, pharmacovigilance aims to detect rare or delayed harms that may not appear in clinical trials. A systematic review of methods for drug safety signal detection using routinely collected observational electronic healthcare data found wide method heterogeneity and no consistently superior approach across drug–event pairs; performance assessment and reporting were heterogeneous, and replicability was limited due to lack of transparency and lack of a gold standard reference set. (18) The review noted growing interest in applying machine learning, alongside epidemiologic designs and adaptations of spontaneous report methods. (18)

Interpretation. For pharmacovigilance, ML may help scale screening, but the evidentiary “hard part” remains causal interpretation: separating signals from confounding and surveillance artefacts, and then translating signals into regulatory or clinical action. I therefore view AI primarily as an augmentation to established pharmacoepidemiology rather than a replacement, with emphasis on reproducible pipelines, auditability, and strong epidemiologic design choices. (18)

Patient messaging, documentation, and LLM-mediated medication text

Medication safety is often determined by text: directions, units, frequency, duration, and contraindications. NLP and LLMs may help detect errors and standardise instructions, but they can also introduce new hazards if they hallucinate, omit, or “smooth over” uncertainty.

A narrative review in *Pharmacotherapy* discussed how NLP can support medication safety tasks across EHR text, literature, reporting systems, and internet-based data sources, highlighting automation potential but also interoperability constraints. (19)

More directly medication-linked, a *Nature Medicine* study evaluated an LLM-based system to prevent medication direction errors in online pharmacies. In benchmark comparisons, two LLM-based baselines recorded 1.51 and 4.38 times more near-miss events than the proposed system, and deployment within a production online pharmacy setting was associated with a reduction in near-miss events reported as 33% (with a reported confidence interval). (20) This is a strong proof-of-concept for a key design principle: **domain logic and safety guardrails can outperform “raw” LLM prompting**, especially when outputs are constrained and evaluated against expert review. (20)

However, even when outputs look fluent, hallucinations and omissions remain central safety concerns. A 2025 framework for assessing clinical safety and hallucination rates of LLMs for medical text summarisation reported a hallucination rate of 1.47% and an omission rate of 3.45% in clinician-annotated evaluations of an LLM document-generation pipeline, while emphasising that iterative workflow refinement can reduce major errors. (21) In medication contexts, even low-frequency hallucinations can be unacceptable if they occur in high-severity categories (e.g., dose, route, insulin type, anticoagulant interactions). (21)

A study assessing ChatGPT-4 on tasks relevant to community pharmacy (drug information, labelling error identification, prescription interpretation, decision support) found that LLM outputs could be satisfactory in certain scenarios, while emphasising the need for rigorous validation across diverse queries, drug classes, populations, and privacy protections. (22)

Interpretation. For patient messaging and documentation, I treat LLMs as “high leverage, high liability” systems. Safe deployment requires constrained tasks, reference grounding, explicit uncertainty, and measurable clinical safety endpoints (not just BLEU scores or subjective “helpfulness”).

Cross-cutting failure modes and why they are predictable

I group medication-AI failure modes into predictable classes:

- **Automation bias and verification failure.** Automation bias—over-reliance on automated advice—has been systematically reviewed, with mitigators including training, emphasising accountability, and design strategies such as confidence information and better presentation of advice. (12) In prescribing contexts, incorrect CDSS can induce omission and commission errors, demonstrating how users can accept false alerts and miss true problems when trust is miscalibrated. (15) A systematic review focusing on automation bias and verification complexity found that such bias can occur even in single-task environments and may relate to cognitive load and task verification difficulty. (13)
- **Dataset shift and performance decay.** Dataset shift—mismatch between development data and deployment data—can cause underperformance and is a major driver of clinical AI malfunction, requiring monitoring and update strategies. (23)
- **Bias and inequity.** An empirical study on a widely used healthcare management algorithm demonstrated racial bias arising from the use of healthcare costs as a proxy for need, resulting in systematically different risk assignment across groups. (24) For medication safety tools, analogous risks include biased detection of ADEs (because documentation/monitoring differs) and biased allocation of pharmacist review resources (because “risk” estimates embed access and utilisation patterns). (24)
- **Methodological fragility and poor reporting.** For prediction models and ML systems, the combination of poor reporting (TRIPOD), the need for structured risk-of-bias assessment (PROBAST), and observed high risk-of-bias prevalence in ML prediction model studies directly threatens safe translation to practice. (9–11)
- **Generative errors (hallucination/omission) and false fluency.** LLM outputs can be clinically unsafe due to hallucinations or omissions even when outputs appear coherent, motivating formal safety assessment frameworks. (21)

These failure modes are not reasons to reject AI; they are reasons to demand **evidence, evaluation discipline, and governance.**

Practical clinician/pharmacist vetting steps before deployment

Below I summarise a pragmatic, clinician-facing vetting framework, anchored to evidence and reporting standards for AI interventions and prediction models. (9–11,25,26) I also incorporate regulatory/governance considerations relevant to medical machine learning (change over time, liability, and lifecycle oversight). (27)

Table: AI applications in medication safety—evidence strength, failure modes, and vetting steps

AI application area	Typical evidence strength (current)	Common failure modes (what goes wrong)	Recommended vetting steps before deployment (clinician/pharmacist-facing)
Prescribing/order support (CDSS/CPOE rules; ML-assisted checks)	Moderate (process outcomes), variable for patient outcomes (2,3,7)	New technology-facilitated errors; unsafe defaults; configuration errors; alert fatigue/overrides; over-trust in alerts (4–7)	Confirm local configuration governance; run “vulnerability testing” using known high-severity scenarios; measure alert burden + override reasons; require monitoring dashboards for new error types; ensure escalation pathways for safety reports (4–7)
Medication reconciliation (eMedRec, NLP extraction, list matching)	Limited-to-moderate (mostly implementation/usability; limited RCT outcome evidence) (14)	Incorrect source reconciliation; missing provenance; duplicate/merged meds; failure at care transitions	Require provenance display for each medication; test against gold-standard histories; validate performance at admission/discharge; assess time-on-task and error recovery; ensure pharmacist review workflow integration (14)
Dosing support (dose calculators; renal dosing alerts; constrained LLM calculation support)	Mixed; strong for narrow rule-based tasks, emerging for constrained LLM calculation studies (15,16)	Automation bias if the tool is wrong; unit conversion errors; unrecognised contraindications/interactions; miscalibration of trust (15,16)	Hard-limit to bounded tasks; force unit confirmation and display intermediate steps; require mandatory human verification for high-alert meds; evaluate automation bias (omission/commission errors) in simulation before go-live (15,16)
ADE prediction (ML risk stratification from EHR)	Emerging; moderate discrimination in meta-analysis, but general high risk-of-bias concerns (9–11,17)	Overfitting, data leakage, label bias; poor calibration; subgroup performance disparities; alert overload without actionability	Demand TRIPOD-level transparency; PROBAST-guided risk-of-bias review; external + temporal validation locally; calibration assessment; subgroup audits; define action bundles and measure clinical outcomes, not only AUC (9–11,17)
Pharmacovigilance/signal detection (ML/NLP on RWD + reports)	Emerging; heterogeneous methods; no consistently superior method; limited replicability (18)	Confounding mistaken as signal; non-reproducible pipelines; poor transparency; shifting coding practices	Require transparent protocols, reproducible pipelines, audit trails; evaluate signal validity with epidemiologic designs; define thresholds for escalation; include post-deployment governance and documentation of model updates (18)
Patient messaging & documentation affecting medication safety (LLMs/NLP for directions, summaries, counselling scripts)	Emerging; promising in constrained domains; safety frameworks still developing (20–22)	Hallucinations and omissions; false fluency; unsafe simplification; privacy risks; over-dependence by staff/patients (21,22)	Constrain outputs; require grounding to authoritative sources; implement “no-guess” policies with explicit uncertainty; run clinical safety error taxonomy assessments; monitor hallucination/omission rates and severity; ensure human review for high-risk outputs (20–22)

Discussion

What works, why it works, and where it breaks

The evidence supports a nuanced position:

1. **AI works best when it is bounded.** Rule-based drug–drug interaction checks, allergy checks, dose calculators, and constrained text extraction tasks have clearer evaluation standards and fit structured safety workflows. (2,3,14,15)
2. **“AI as prediction” is not the same as “AI as prevention.”** ADE prediction models with modest-to-moderate discrimination (e.g., pooled AUC ~0.72) may still fail to reduce harm if they trigger unhelpful alerts or if predicted risk is not linked to an effective preventive action set. (17)
3. **Human factors can dominate statistical performance.** High alert override rates and automation bias demonstrate that clinician behaviour is integral to safety performance. (6,12,13,15)
4. **Generative AI introduces new categories of catastrophic error.** Hallucinations and omissions in LLM outputs are qualitatively different from typical CDSS false positives: the system may generate plausible, novel clinical statements, creating a strong risk of “confidently wrong” medication instructions unless outputs are constrained, grounded, and safety-audited. (20,21)
5. **Bias and dataset shift are not edge cases.** They are expected properties of clinical deployments. Dataset shift can occur with changing populations and workflows; biased proxy variables can systematically distort risk assignment and resource allocation. (23,24)

How I would operationalise “vetting” in real medication workflows

In medication safety, clinicians and pharmacists need a **minimum viable governance standard** that can be applied before patient exposure. Based on the evidence and reporting expectations, I would operationalise vetting in three layers:

Layer 1: Evidence and intent (pre-procurement).

I would require a precise description of intended use, user role, and actionability, and I would consider it a red flag if the vendor cannot specify what the tool changes in workflow and what “success” looks like. For AI interventions evaluated in trials, completeness and clarity expectations are codified by CONSORT-AI and SPIRIT-AI: the AI intervention, its integration into setting, human–AI interaction, and handling of inputs/outputs must be explicit. (25,26)

Layer 2: Local verification (pre-deployment).

For prediction models, I would require at minimum TRIPOD-aligned disclosure and a PROBAST-informed risk-of-bias assessment. (9,10) Given the high observed prevalence of high risk-of-bias in ML prediction models, I would not accept “AUC in the paper” as sufficient. (11) I would also require temporal validation (recent data) and subgroup performance checks to reduce the risk of silent inequities. (23,24)

Layer 3: Lifecycle governance (post-deployment monitoring).

Unlike static software, medical ML may drift (dataset shift), and governance must include monitoring and change control. (23,27) At minimum, I would mandate: drift monitoring dashboards, thresholds for “suspend/caution” states, incident reporting mechanisms tied to medication safety committees, and periodic re-validation.

Regulatory and governance considerations

From a governance standpoint, AI blurs boundaries between clinical practice, software, and medical devices. A legal and policy analysis of regulatory responses to medical machine learning emphasises challenges in evaluating safety and effectiveness for ML-based medical devices and highlights that regulatory approaches must address the possibility of change over time. (27) Complementing this, an analysis of the FDA-approved/cleared AI/ML-enabled medical device landscape reported 343 devices compiled by the FDA by June 2021, with most cleared via the 510(k) pathway and a high proportion functioning as software as a medical device (SaMD). (28) Although most of these devices are in radiology/cardiology rather than medication safety per se, the lifecycle governance issues (intended use, generalisability, update

pathways) are directly relevant to medication AI tools that may be classified as SaMD or used as clinical decision support. (27,28)

India-facing considerations (transferability and implementation)

In India, medication safety initiatives often operate within resource constraints and heterogeneous digital maturity across settings. This amplifies several issues identified in the evidence: data quality variation (affecting ML), alert burden with limited staffing, multilingual documentation needs (affecting NLP/LLMs), and uneven access to calibration populations (affecting equity). The most transferable strategy appears to be **bounded tools with high transparency** (dose calculators, structured checks, reconciliation provenance), combined with human review where the marginal safety value is highest (high-alert medications, paediatrics, anticoagulants). This aligns with the evidence that technology effects are heterogeneous and implementation-determined, and that errors can be shifted rather than eliminated without strong governance. (2,4,5,23)

Limitations

This manuscript is a narrative synthesis rather than a full systematic review with duplicate screening and formal meta-analysis across all domains; as a result, I may under-represent some subliteratures, particularly outside PubMed-indexed journals or outside high-income settings. The evidence base is also heterogeneous in outcome metrics: CDSS studies often report process measures and error types, prediction model studies report discrimination metrics (e.g., AUC) that may not translate to clinical benefit, and LLM studies vary widely in task framing and safety endpoints. (7,9–11,17,21) Finally, some cited LLM studies are rapidly emerging, and their findings may not generalise across models, prompts, or institutional contexts.

Conclusion

Medication safety is an unusually demanding proving ground for AI because it combines high stakes, frequent decisions, workflow pressure, and complex causality. The mature evidence on CPOE/CDSS supports cautious benefit with heterogeneous outcomes and persistent unintended consequences. (2–7) The newer evidence for ML prediction (e.g., ADE prediction with pooled AUC ~0.72) suggests potential but also underscores the need for rigorous reporting, external validation, and bias control. (9–11,17) LLMs may reduce specific, constrained medication text errors when designed with domain logic and guardrails, but hallucinations and omissions require formal safety measurement and strong human oversight. (20,21)

Before any AI tool touches patient care in medication workflows, I recommend a minimum safety standard: clearly bounded intended use; transparent reporting consistent with AI trial guidance where applicable; prediction-model standards (TRIPOD/PROBAST) where applicable; local external/temporal validation; explicit measurement and mitigation of automation bias, dataset shift, and inequity; and a post-deployment monitoring plan with clear stop rules and accountability. (9–13,23–27)

References

1. Bates DW, Cullen DJ, Laird N, et al; ADE Prevention Study Group. Incidence of adverse drug events and potential adverse drug events: implications for prevention. *JAMA*. 1995;274(1):29-34. PMID:7791255.
2. Nuckols TK, Smith-Spangler C, Morton SC, et al. The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. *Syst Rev*. 2014;3:56. doi:10.1186/2046-4053-3-56. PMID:24894078.
3. Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*. 2005;293(10):1223-1238. doi:10.1001/jama.293.10.1223. PMID:15755945.
4. Koppel R, Metlay JP, Cohen A, et al. Role of computerized physician order entry systems in facilitating medication errors. *JAMA*. 2005;293(10):1197-1203. doi:10.1001/jama.293.10.1197. PMID:15755942.
5. Schiff GD, Amato MG, Egale T, et al. Computerised physician order entry-related medication errors: analysis of reported errors and vulnerability testing of current systems. *BMJ Qual Saf*. 2015;24(4):264-271. doi:10.1136/bmjqs-2014-003555. PMID:25595599.
6. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. *J Am Med Inform Assoc*. 2006;13(2):138-147. doi:10.1197/jamia.M1809. PMID:16357358.

7. Page N, Baysari MT, Westbrook JI. A systematic review of the effectiveness of interruptive medication prescribing alerts in hospital CPOE systems to change prescriber behavior and improve patient safety. *Int J Med Inform.* 2017;105:22-30. doi:10.1016/j.ijmedinf.2017.05.011. PMID:28750908.
8. Shahmoradi L, Safdari R, Ahmadi H, Zahmatkeshan M. Clinical decision support systems-based interventions to improve medication outcomes: a systematic literature review on features and effects. *Med J Islam Repub Iran.* 2021;35:27. doi:10.47176/mjiri.35.27. PMID:34169039.
9. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* 2015;162(1):55-63. doi:10.7326/M14-0697. PMID:25560714.
10. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med.* 2019;170(1):51-58. doi:10.7326/M18-1376. PMID:30596875.
11. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ.* 2021;375:n2281. doi:10.1136/bmj.n2281. PMID:34670780.
12. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc.* 2012;19(1):121-127. doi:10.1136/amiajnl-2011-000089. PMID:21685142.
13. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc.* 2017;24(2):423-431. doi:10.1093/jamia/ocw105. PMID:27516495.
14. Marien S, Krug B, Spinewine A. Electronic tools to support medication reconciliation: a systematic review. *J Am Med Inform Assoc.* 2017;24(1):227-240. doi:10.1093/jamia/ocw068. PMID:27301747.
15. Lyell D, Magrabi F, Raban MZ, et al. Automation bias in electronic prescribing. *BMC Med Inform Decis Mak.* 2017;17(1):28. doi:10.1186/s12911-017-0425-5. PMID:28302112.
16. Levin C, Orkaby B, Kerner E, Saban M. Can large language models assist with pediatric dosing accuracy? *Pediatr Res.* 2025;98(5):1760-1765. doi:10.1038/s41390-025-03980-8. PMID:40057653.
17. Hu Q, Li J, Li X, Zou D, Xu T, He Z. Machine learning to predict adverse drug events based on electronic health records: a systematic review and meta-analysis. *J Int Med Res.* 2024;52(12):3000605241302304. doi:10.1177/03000605241302304. PMID:39668733.
18. Coste A, Wong A, Bokern M, Bate A, Douglas IJ. Methods for drug safety signal detection using routinely collected observational electronic health care data: A systematic review. *Pharmacoepidemiol Drug Saf.* 2023;32(1):28-43. doi:10.1002/pds.5548. PMID:36218170.
19. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. *Pharmacotherapy.* 2018;38(8):822-841. doi:10.1002/phar.2151. PMID:29884988.
20. Pais C, Liu J, Voigt R, et al. Large language models for preventing medication direction errors in online pharmacies. *Nat Med.* 2024;30(6):1574-1582. doi:10.1038/s41591-024-02933-8. PMID:38664535.
21. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med.* 2025;8(1):274. doi:10.1038/s41746-025-01670-7. PMID:40360677.
22. Shin E, Hartman M, Ramanathan M. Performance of the ChatGPT large language model for decision support in community pharmacy. *Br J Clin Pharmacol.* 2024;90(12):3320-3333. doi:10.1111/bcp.16215. PMID:39191671.
23. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med.* 2021;385(3):283-286. doi:10.1056/NEJMc2104626. PMID:34260843.
24. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453. doi:10.1126/science.aax2342. PMID:31649194.
25. 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):1364-1374. doi:10.1038/s41591-020-1034-x. PMID:32908283.
26. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health.* 2020;2(10):e549-e560. doi:10.1016/S2589-7500(20)30219-3. PMID:33328049.
27. Minssen T, Gerke S, Aboy M, Price WN II, Cohen G. Regulatory responses to medical machine learning. *J Law Biosci.* 2020;7(1):lsaa002. doi:10.1093/jlb/lsaa002. PMID:34221415.