Original Article

Counterfactual Medication Planning before Treatment Is Changed

Noah Williams[1]*, Grace Collins[1], Ethan Brooks[2], Daniel Green[3]

[1]Department of Counterfactual Reasoning in Pharmacy, Faculty of Pharmacy, University of Sydney, Sydney, Australia. [2]Department of Treatment Change Planning and AI, Faculty of Pharmacy, University of Queensland, Brisbane, Australia. [3]Department of Medication Scenario Simulation, Faculty of Pharmacy, University of New South Wales, Sydney, Australia.

Abstract

Medication decisions are commonly informed by predicted risks, expected responses, previous treatment experience, and population-level evidence. Yet the practical question faced before treatment is changed is not merely whether an adverse event or therapeutic failure is likely. It is how the patient's future might differ under continuation, dose modification, switching, augmentation, tapering, discontinuation, delayed action, or additional information collection. Current predictive, causal, medication-safety, and clinical decision-support approaches address parts of this problem but do not provide an integrated structure for comparing alternative medication trajectories while preserving uncertainty, professional responsibility, and patient priorities. This article proposes a Counterfactual Medication-Planning Framework for organizing treatment-change deliberation. The framework begins with an explicitly defined decision time and patient–medication state, identifies feasible treatment strategies, specifies outcomes and time horizons, and constructs parallel multidimensional trajectories. Each trajectory is qualified through an assumption and confounding ledger and an uncertainty envelope. An identifiability and comparability gate determines whether alternatives may proceed to human deliberation, require additional information or restriction, or should not be compared. The framework separates trajectory generation from medication authorization and places pharmacists, prescribers, patients, and accountable organizations within a bounded decision and monitoring process. Evaluation would require causal, technical, medication-safety, human-factors, equity, transportability, implementation, and lifecycle-governance evidence. Important boundaries include unmeasured confounding, incomplete longitudinal records, poor treatment overlap, uncertain adherence, unstable patient preferences, rare or delayed harms, and the impossibility of observing both treatment futures in the same patient. The contribution is an original non-empirical synthesis intended to make the assumptions, uncertainties, responsibilities, and abstention conditions of pre-change medication planning explicit. It requires empirical validation before clinical use.

Keywords: Digital pharmacy, Artificial intelligence, Pharmacy practice, Medication safety, Human–AI collaboration, Clinical decision support

Introduction

Clinical prediction often estimates what is likely to occur given the patient's observed characteristics and current care. A medication-change decision asks a different question: how might the patient's future differ if treatment were continued, modified, replaced, supplemented, reduced, or stopped? Predicting an outcome under observed practice does not automatically estimate outcomes under alternative interventions, because treatment assignment, clinical judgment, disease severity, adherence, monitoring, and access may influence both the treatment received and the outcome observed. Causal machine-learning methods formalize this distinction through potential outcomes and treatment contrasts, but their validity remains conditional on the data-generating process and causal assumptions [1].

This distinction is especially important in longitudinal pharmacotherapy. Medication decisions occur within evolving clinical states rather than isolated encounters. Previous response, cumulative exposure, adverse effects, concurrent medicines, organ function, adherence, patient goals, and previous treatment changes may affect both the next intervention and subsequent outcomes. Observational clinical data can support causal analyses, but reviews of practice show recurrent weaknesses in design specification,

Address for correspondence: Noah Williams, Department of Counterfactual Reasoning in Pharmacy, Faculty of Pharmacy, University of Sydney, Sydney, Australia.
noah.williams@sydney.edu.au
Received: 28 February 2026; **Accepted:** 03 May 2026

How to cite this article: Williams N, Collins G, Brooks E, Green D. Counterfactual Medication Planning before Treatment Is Changed. Arch Pharm Pract. 2026;17(2):47-56. https://doi.org/10.51847/Qqs5hU4vV0

confounding control, temporal alignment, missing-data handling, and reporting [2]. A numerically precise output may therefore remain poorly suited to treatment comparison if the underlying intervention, comparator, follow-up period, or assumptions are unclear.

Counterfactual clinical-prediction models offer one route to individualized treatment-effect estimation by predicting outcomes under specified alternatives. Their use in randomized-trial data illustrates how treatment interactions and potential outcomes may be represented at the individual level, while also showing that model form, calibration, trial population, outcome definition, and validation strategy determine what the estimates can support [3]. Such models do not reveal two observed futures for one patient. They estimate an unobserved alternative using evidence and assumptions, and the resulting contrast should be interpreted accordingly.

The proposed article therefore treats counterfactual medication planning as broader than generating a treatment-effect number. Actionable healthcare questions require causal reasoning, but they also require clinically feasible alternatives, medication-specific safety interpretation, uncertainty communication, human deliberation, and governance [4]. A treatment might appear favorable for one endpoint while increasing monitoring burden, interaction risk, adherence difficulty, cost, or another patient-valued harm. Conversely, continuation may avoid transition risks but prolong ineffective or burdensome treatment. The relevant object is therefore a set of qualified future trajectories rather than a single risk score.

The central proposition is that a medication should not be changed solely because one model ranks an alternative above current treatment. Before a comparison can influence care, the strategies, outcomes, time horizon, data provenance, causal assumptions, uncertainty, patient priorities, and professional responsibilities must be explicit. The proposed framework integrates these elements while preserving a distinction between evidence-supported methodological components and the original architecture through which this article relates them.

Counterfactual Reasoning in Pharmacotherapy

Counterfactual reasoning concerns contrasts between outcomes expected under different specified interventions. In pharmacotherapy, the alternatives must be defined as treatment strategies rather than vague labels. “Continue treatment” may include dose, schedule, adherence expectations, monitoring, rescue therapy, and rules for subsequent modification. “Switch treatment” must identify the replacement medicine, transition process, washout or overlap, expected adherence, and management of intercurrent events. Without this specificity, apparently different trajectories may represent inconsistent or clinically incomparable treatment versions.

A second distinction concerns baseline risk and treatment-effect heterogeneity. Patients at high risk of an outcome do not necessarily receive the greatest benefit from an intervention, and variables predicting prognosis may not predict differential treatment response. The Predictive Approaches to Treatment effect Heterogeneity statement emphasizes prespecification, transparent modeling, and separation of risk-based and effect-based analyses [5]. Individualized treatment-effect models must also be assessed for whether they discriminate between patients with different relative effects and whether their predicted treatment contrasts are calibrated [6]. These assessments remain separate from causal identification: a well-calibrated model within one dataset can still be biased by inappropriate comparison or unmeasured confounding.

Pharmacotherapy also contains repeated decisions. Treatment may be intensified after inadequate response, reduced after toxicity, changed after laboratory deterioration, or maintained while further evidence is collected. Dynamic treatment regimes represent sequences of decision rules conditioned on evolving patient information. Interpretable regimes demonstrate how such sequences can be expressed in clinically comprehensible terms, but exploratory rules derived from data should not be treated automatically as treatment instructions [7]. The relevant contribution to this article is the representation of medication planning as a time-dependent process, not an endorsement of any particular algorithm.

Real-world patient data may expand the range of populations and treatment patterns available for study, but individualized treatment-effect estimation from those data faces confounding, inadequate overlap, missingness, measurement error, treatment-selection bias, and representation problems. These challenges are particularly important in clinical pharmacology, where dose, exposure, comorbidity, co-medication, adherence, and monitoring may vary over time [8]. Machine learning can model complex relationships, but flexibility does not remove the need for causal design or external validation.

A treatment-effect model is not decision-ready merely because heterogeneity is estimated; its estimand, subgroup strategy, discrimination, calibration, and validation context must be explicit [5, 6]. Counterfactual medication planning is consequently proposed as a deliberative structure surrounding, rather than replacing, effect estimation. It asks whether the alternatives are clinically feasible, whether their trajectories are causally interpretable, which outcomes are omitted, how uncertainty should limit reliance, and who is responsible for reviewing the comparison. The framework may incorporate trial evidence, observational analyses, mechanistic knowledge, professional judgment, and patient preferences, but none is sufficient alone.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for

definitions, components, relationships, and intended uses for the article Counterfactual Medication Planning before Treatment Is Changed.

Table 1. Definitions, components, relationships, and intended uses for the article Counterfactual Medication Planning before Treatment Is Changed.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement
Decision time	Medication questions may be framed without a defined point at which alternatives are compared.	Current treatment, clinical state, reason for reconsideration, available evidence	Proposed: anchor all information and alternatives to one explicit pre-change decision point.	Prevents future information from being incorporated as though it were available at the original decision.	Temporal data provenance and clinically meaningful decision definition [2].	Look-ahead bias, misaligned measurements, retrospective reinterpretation	Defining the time point does not establish causal comparability.
Patient–medication state	A diagnosis or risk score may omit determinants of treatment feasibility and response.	Medicines, doses, indications, response, harms, adherence, laboratory data, comorbidities, preferences	Proposed: construct a time-indexed state containing decision-relevant medication and patient information.	Makes the factual starting point of each alternative explicit.	Longitudinal completeness, measurement validity, and relevance to the decision [1].	Missing adherence, unrecorded nonprescription use, inaccurate medication history	The recorded state remains an incomplete representation of the patient.
Feasible alternative set	Models may compare interventions that are unavailable, unacceptable, or clinically incoherent.	Treatment options, contraindications, formulary, access, patient goals, professional judgment	Proposed: restrict comparison to operationally and clinically plausible strategies.	Aligns modeled alternatives with decisions that could actually be considered.	Explicit intervention and comparator definitions [3].	Positivity failure, unrealistic treatment versions, omitted viable alternatives	Feasibility does not imply that an alternative is appropriate.
Estimand	"Treatment effect" may be stated without defining whose outcome, under which strategies, over what period.	Alternatives, outcome, population or patient context, horizon, intercurrent events	Evidence-supported causal specification connected to the proposed planning process.	Clarifies the exact contrast being estimated.	Transparent heterogeneity and effect-estimation strategy [5].	Ambiguous comparator, incompatible outcomes, post-treatment conditioning	A precise estimand may still be unidentified.
Counterfactual trajectory	Single endpoints may conceal timing, competing outcomes, and evolving treatment consequences.	Longitudinal state, treatment strategy, outcome processes, follow-up horizon	Proposed: represent each alternative as a multidimensional path rather than one scalar prediction.	Supports comparison of benefit, harm, burden, and monitoring over time.	Valid treatment-effect modeling and calibrated evaluation [6].	Trajectory instability, extrapolation, omitted delayed harm	A modeled trajectory is not the observed future of the patient.
Dynamic decision rule	Treatment may change in response to evolving clinical information.	Time-varying measurements, response, toxicity, adherence, thresholds for reconsideration	Evidence-supported sequential structure incorporated into the proposed trajectory model.	Represents adaptation rather than fixed one-time exposure.	Interpretable and clinically defensible sequential strategy specification [7].	Policy instability, inappropriate thresholds, feedback bias	An estimated rule is hypothesis-generating until prospectively evaluated.
Assumption ledger	Causal assumptions are often hidden behind model output.	Exchangeability, consistency, positivity, measurement, interference, transportability	Proposed: display assumptions beside each comparison and record their evidentiary status.	Makes causal limitations reviewable by clinical and technical users.	Real-world treatment-effect methodology and sensitivity analysis [8].	False confidence, inconsistent assumptions across alternatives	Documentation of an assumption does not make it true.
Uncertainty envelope	Conventional predictive intervals may omit causal and	Sampling variation, model uncertainty, measurement error, confounding,	Proposed: separate uncertainty classes and prevent their	Supports bounded interpretation and abstention.	Calibration and causal-validity evidence [6].	Narrow intervals despite weak identification,	Uncertainty communication cannot repair nonidentifiability.

	contextual uncertainty.	transportability, preference stability	collapse into one confidence score.			misleading precision	
Human deliberation	Model ranking may be mistaken for treatment authorization.	Trajectory comparison, safety considerations, patient values, professional roles	Proposed: require pharmacist–prescriber–patient interpretation before a disposition is formed.	Preserves clinical accountability and preference-sensitive reasoning.	Evidence of comprehensibility, workflow fit, and appropriate reliance	Automation bias, role ambiguity, excessive deference or rejection	Human review does not itself validate an unreliable comparison.
Monitoring loop	Postchange follow-up may be treated as proof that the selected counterfactual was correct.	Observed response, harms, adherence, incidents, new information	Proposed: use follow-up to update the factual state and governance record.	Supports learning, reassessment, and incident response.	Prospective outcome and implementation evaluation	Delayed harms, missing follow-up, attribution errors	The unchosen patient-specific trajectory remains unobserved.

Proposed Counterfactual Medication-Planning Framework

The Counterfactual Medication-Planning Framework is proposed as a nine-stage architecture for organizing the reasoning that should precede treatment change. It is not a treatment-selection algorithm. Its purpose is to make visible what is being compared, how alternative futures are generated, which assumptions support them, where comparison becomes unreliable, and how model output relates to professional and patient deliberation.

The first stage frames the medication question at a defined decision time. The second represents the current patient–medication state. The third identifies feasible strategies, which may include continuation, adjustment, switching, augmentation, tapering or discontinuation, temporary deferment, or targeted information acquisition. The fourth specifies the estimand, outcomes, horizon, and handling of intercurrent events. Work connecting causal inference and digital-twin concepts supports the need to define interventions, estimands, heterogeneity, and transportability before simulated alternatives are interpreted as counterfactual trajectories [9].

The fifth stage constructs parallel trajectory representations. Longitudinal modeling can jointly represent evolving clinical measurements and event risks rather than treating future outcomes as independent point predictions [10]. Patient-trajectory research similarly emphasizes the connection among problem formulation, temporal data processing, modeling, evaluation, and interpretation [11]. In the proposed framework, each trajectory contains expected therapeutic benefit, adverse effects, treatment burden, adherence feasibility, monitoring needs, and patient-valued consequences. These dimensions are not collapsed automatically into one utility score because their relative importance may be uncertain, context-dependent, or preference-sensitive.

The sixth stage applies an assumption and confounding ledger, while the seventh applies an uncertainty envelope and an identifiability and comparability gate. The gate does not ask whether one trajectory has the highest model score. It asks whether the alternatives are defined consistently, supported by relevant evidence, sufficiently overlapping, and qualified by uncertainty for bounded deliberation. Its proposed dispositions are: proceed to deliberation, proceed conditionally, restrict the comparison, obtain additional information, defer the decision, or abstain from counterfactual ranking.

The eighth stage is human deliberation. Medication-specific decision-support research shows that machine learning may assist a bounded task such as prioritizing prescriptions for pharmacist review, but task-level performance does not establish patient benefit or confer authority to change therapy [12]. The framework therefore separates trajectory production from treatment authorization. Pharmacists may contribute medication-history verification, interaction and adverse-effect interpretation, adherence assessment, and monitoring design. Prescribers retain responsibility for diagnosis, treatment authorization, and integration with the wider care plan. Patients contribute goals, lived experience, burden tolerance, and preferences that cannot be inferred reliably from the record alone.

The ninth stage records a bounded decision disposition and establishes monitoring. Possible dispositions include considering a treatment alternative, temporarily maintaining treatment, gathering more information, escalating review, or abstaining from comparison. Monitoring updates the factual patient state and may identify incidents, drift, or new constraints. It does not reveal what would have happened under the unchosen strategy. The overall architecture is an original synthesis: the literature supports its component concepts, but not their exact integration, ordering, gate logic, or clinical effectiveness.

Figure 1 presents the original conceptual synthesis for counterfactual medication-planning framework, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

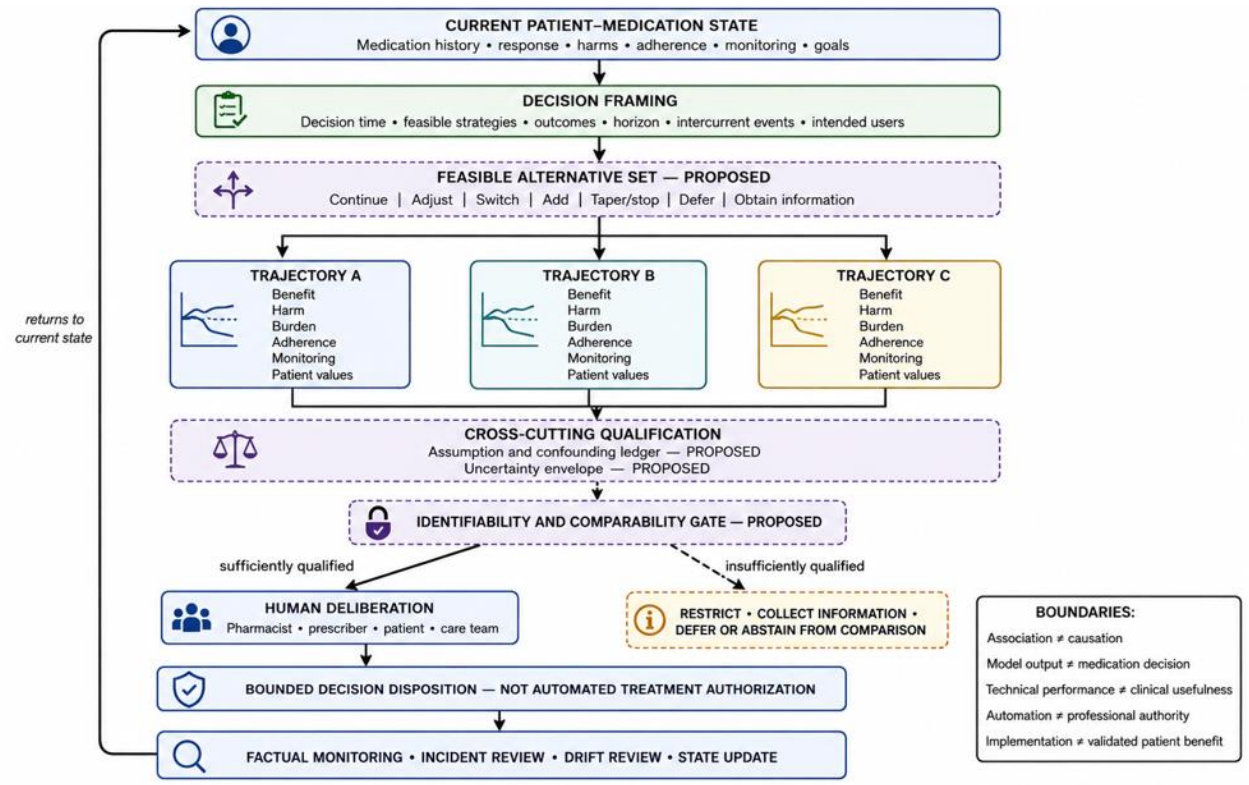

Figure 1. Counterfactual Medication-Planning Framework

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for evaluation criteria, evidence requirements, and failure modes for the article Counterfactual Medication Planning before Treatment Is Changed.

Table 2. Evaluation criteria, evidence requirements, and failure modes for the article Counterfactual Medication Planning before Treatment Is Changed.

Architecture layer or process stage	Core function	Information flow	Interaction with other components	Evaluation criterion	Potential failure mode	Human or organizational responsibility	Readiness boundary
Decision framing	Define the question before alternatives are modeled.	Clinical concern and current treatment enter; explicit strategies, outcomes, and horizon leave.	Determines what the state representation and trajectory engine must contain.	Completeness and clinical coherence of the treatment contrast.	Vague comparator, unsuitable outcome, post hoc horizon	Pharmacist and prescriber verify question; patient confirms relevance.	Proposed: do not construct trajectories until the contrast is sufficiently specified.
Patient–medication state	Represent the factual baseline available at decision time.	Medication, clinical, monitoring, adherence, access, and preference data are assembled.	Supplies initial conditions to every alternative trajectory.	Temporal validity, completeness, provenance, and relevance.	Medication-reconciliation error, missing adherence, delayed data	Pharmacy and clinical teams verify state; organization ensures data provenance.	Incomplete state requires restriction or information acquisition.
Feasible-alternative layer	Exclude impossible or incoherent treatment strategies.	Candidate strategies are screened against contraindications, access, preferences, and care context.	Defines the branches that proceed to trajectory construction.	Clinical feasibility and consistency of treatment versions.	Comparing unavailable medicines or unrealistic treatment transitions	Prescriber authorizes clinical plausibility; pharmacist assesses medication feasibility.	Feasibility permits comparison, not treatment selection.
Estimand layer	Specify the causal contrast.	Strategies, outcome vector, horizon, population context, and intercurrent events are formalized.	Connects decision framing to evidence and trajectory outputs.	Alignment between clinical question, data, and estimand [9].	Estimand mismatch, conditioning on post-treatment variables	Clinical and causal-method experts review specification.	No causal interpretation without an explicit estimand.

Trajectory-construction layer	Generate multidimensional alternative futures.	The common patient state is propagated under each strategy.	Produces benefit, harm, burden, adherence, and monitoring paths.	Temporal coherence, calibration, reproducibility, and baseline comparison [10].	Implausible extrapolation, omitted outcome, unstable long-horizon projection	Technical team validates modeling; clinicians review plausibility.	Trajectories remain analytic objects rather than recommendations.
Patient-trajectory interpretation layer	Translate temporal patterns into clinically meaningful representations.	Events and measurements are organized into understandable paths.	Links model outputs to professional deliberation.	Clinical interpretability and preservation of relevant longitudinal information [11].	Oversimplified chronology, missing context, visual overload	Human-factors and clinical teams assess comprehension.	Interpretability cannot compensate for causal invalidity.
Medication-safety layer	Identify medication-specific errors and foreseeable harms.	Prescription, interaction, contraindication, monitoring, and transition risks enter the review.	Constrains trajectory comparison and deliberation.	Sensitivity to clinically important medication risks and manageable review burden [12].	False reassurance, excessive alerts, omitted rare harm	Pharmacists lead safety review; organization defines escalation routes.	Task performance does not authorize medication change.
Assumption and confounding ledger	Expose causal and measurement conditions.	Assumptions are attached to each strategy and evidence source.	Qualifies every trajectory and feeds the comparability gate.	Transparency, consistency, sensitivity analysis, and reviewability.	Hidden confounding, unsupported overlap, ambiguous exposure	Causal-method and clinical reviewers document and challenge assumptions.	Unresolved critical assumptions may require abstention.
Uncertainty envelope	Separate distinct sources of uncertainty.	Predictive, measurement, causal, transportability, and preference uncertainty accompany outputs.	Prevents one confidence value from obscuring identification limitations.	Calibration, coverage, sensitivity, and understandable communication.	Misleading precision or conflation of statistical and causal uncertainty	Technical team quantifies; clinicians communicate; governance sets limits.	Narrow uncertainty does not override failed identification.
Identifiability and comparability gate	Decide whether alternatives are fit for bounded deliberation.	Qualified trajectories enter; proceed, restrict, defer, collect information, or abstain leaves.	Controls access to human deliberation.	Explicit and reproducible gate rationale.	Automatic passage based on model score, inconsistent gate application	Multidisciplinary accountable review body.	Proposed: no trajectory ranking when the comparison is materially nonidentifiable.
Human-deliberation layer	Integrate evidence, safety, feasibility, and patient priorities.	Qualified comparison is interpreted by pharmacist, prescriber, patient, and care team.	Converts analytic output into a bounded disposition.	Comprehension, calibrated reliance, workload, role clarity, and preference integration.	Automation bias, diffusion of responsibility, tokenistic patient involvement	Named clinical decision-maker; documented pharmacist and patient contributions.	Deliberation is required but does not validate the model.
Monitoring and governance layer	Observe the factual trajectory and manage change.	Outcomes, harms, incidents, new data, and drift update the record.	Feeds the next decision cycle and organizational oversight.	Timely monitoring, incident review, auditability, restriction, and requalification.	Silent drift, unreviewed updates, delayed detection of harm	Organization assigns model, medication-safety, and clinical owners.	Implementation remains conditional on continuing local evidence and oversight.

Constructing Alternative Treatment Trajectories

An alternative treatment trajectory should begin with a strategy definition rather than a medication name alone. Continuing therapy may mean continuing the same dose unless toxicity occurs, whereas switching may involve tapering, overlap, washout, intensified monitoring, or rescue treatment. Per-protocol causal analyses demonstrate why sustained strategies, adherence, treatment changes, censoring, and time-varying conditions must be defined when outcomes are attributed to a treatment course [13]. In medication planning, these elements should be specified prospectively within the conceptual comparison, even when the available evidence cannot estimate every component reliably.

The trajectory object is proposed as a linked set of clinically meaningful processes. It should include expected therapeutic response, adverse effects, drug–drug and drug–disease concerns, treatment burden, adherence feasibility, monitoring

requirements, and outcomes important to the patient. Each process may follow a different time scale. Symptom benefit may emerge quickly, laboratory toxicity may develop later, withdrawal effects may occur during transition, and long-term complications may remain poorly represented in available data. A trajectory should therefore retain timing and uncertainty instead of combining all consequences immediately into one score.

Sequential policy-learning research illustrates how treatment decisions can be related to evolving patient states. The Artificial Intelligence Clinician study derived time-dependent treatment strategies from retrospective intensive-care data, demonstrating the technical possibility of estimating policies over a clinical trajectory [14]. That example does not establish that retrospectively derived policies are safe or appropriate for medication planning in other settings. It instead shows why the state, action, timing, reward definition, and observed clinician behavior must be examined before a learned strategy is interpreted.

Healthcare reinforcement-learning guidance identifies additional risks, including confounding by indication, incomplete state representation, inappropriate reward functions, unstable policy evaluation, and unsafe extrapolation [15]. These concerns apply even when the proposed framework does not use reinforcement learning. Any method that ranks longitudinal strategies may favor actions that appear beneficial because of historical treatment-selection patterns rather than causal effects. Consequently, trajectory construction should be accompanied by defensible comparators, off-policy or external evaluation where relevant, safety constraints, and an explicit abstention route.

Target-trial emulation provides a complementary design logic. It requires articulation of eligibility, treatment strategies, assignment procedures, follow-up, outcomes, causal contrasts, and analysis. Federated target-trial emulation further shows that treatment-effect estimation may be coordinated across distributed data while preserving local data control, although site heterogeneity and data compatibility remain important [16]. For counterfactual medication planning, the relevant principle is not that every decision requires a formal federated study. It is that the target comparison should be designed before analytic estimation and should not be reconstructed opportunistically from whichever data are most convenient.

Alternative trajectories may be assembled from multiple evidence sources, but their contributions must remain distinguishable. Randomized evidence may support average comparative effects; observational evidence may inform underrepresented populations or longer-term outcomes; pharmacologic knowledge may constrain implausible paths; and patient reports may reveal burdens absent from structured data. The framework does not propose that these sources be blended into an unqualified composite. Their designs, populations, outcomes, and assumptions should remain visible within the trajectory record.

The proposed construction process ends with a trajectory dossier for each feasible strategy. The dossier contains the intervention definition, outcome vector, time horizon, data sources, causal estimand, assumptions, uncertainty classes, known evidence gaps, and conditions that would invalidate or restrict the comparison. It should also state which outcomes have not been modeled. This negative specification is important because omitted harms or burdens can disappear from a visual comparison despite remaining clinically relevant.

A counterfactual trajectory is therefore an evidence-qualified representation of a possible future, not an observation of the future that would have occurred. Its role is to discipline deliberation by making alternatives, assumptions, and missing information explicit. It should not convert uncertainty into false precision or substitute model ranking for a medication decision.

Assumptions, Confounding, Uncertainty, and Nonidentifiability

Counterfactual medication planning depends on assumptions that cannot be confirmed solely by model performance. Exchangeability requires that relevant determinants of treatment selection and outcome are measured adequately. Consistency requires each treatment strategy to correspond to a sufficiently defined intervention. Positivity requires patients with comparable clinical states to have realistic opportunities to receive the alternatives being compared. Temporal ordering, exposure measurement, adherence, interference, and outcome ascertainment also affect whether an estimated contrast can be interpreted causally.

Unmeasured confounding is particularly consequential when treatment changes are prompted by deterioration, toxicity, clinician concern, or patient preference. Sensitivity analysis can estimate how strongly an omitted variable would need to relate to treatment and outcome to alter an estimated association. Such analysis makes dependence on assumptions more interpretable, but it does not demonstrate that unmeasured confounding is absent [17]. The proposed assumption ledger should therefore record both the analytical adjustment performed and the clinically plausible factors that remain unavailable or poorly measured.

Individualized treatment-effect estimates require uncertainty representations that respond to weakened causal assumptions. Robust conformal approaches illustrate how intervals for individual effects may widen when hidden confounding is permitted [18]. This principle supports a distinction between a narrow prediction interval and a well-supported treatment comparison. Precision conditional on a strong assumption should not be displayed as unconditional certainty.

Observational effect estimates may also contain systematic error not represented by conventional confidence intervals. Empirical calibration using control outcomes demonstrates that nominal intervals can substantially understate uncertainty when residual biases remain [19]. Although population-level calibration does not directly solve individual-effect uncertainty, it reinforces the need to treat sampling variation as only one component of the uncertainty surrounding medication trajectories.

The framework consequently proposes an uncertainty envelope with six distinguishable domains: random outcome variability, model uncertainty, measurement uncertainty, causal-identification uncertainty, transportability uncertainty, and preference uncertainty. These domains should not be collapsed into one confidence score. A model may be statistically stable while the treatment contrast remains unidentified; conversely, a causally defensible average contrast may remain uncertain for a particular patient because relevant effect modifiers or preferences are poorly known.

The identifiability and comparability gate is proposed to convert these limitations into explicit dispositions. A comparison may proceed when strategies are coherent and assumptions are sufficiently defensible; proceed conditionally when uncertainty can be communicated and managed; be restricted to a narrower population, outcome, or horizon; trigger collection of additional information; or be deferred. Where material confounding, absent overlap, inconsistent treatment versions, or missing outcomes prevent a defensible contrast, the framework requires abstention from trajectory ranking.

Nonidentifiability is not equivalent to indecision. It is a substantive finding that available evidence cannot distinguish the alternatives adequately for the proposed use. Sensitivity analysis can reveal dependence on assumptions, but it cannot convert an unsupported comparison into an identified causal effect.

Evaluation and Safe Decision Integration

Evaluation should begin by separating model performance from performance of the complete medication-planning process. A trajectory model may be reproducible and calibrated yet fail because relevant medicines are missing, outputs arrive too late, professionals misunderstand uncertainty, patients cannot express preferences, or responsibility for acting on the comparison is unclear. Early clinical evaluation of artificial-intelligence decision support therefore requires examination of intended users, workflow, safety, human factors, and clinical context rather than technical accuracy alone [20].

Medication-specific evaluation must also distinguish system activity from meaningful benefit. Electronic drug–drug interaction alerts may alter clinician behavior, but evidence of patient-important benefit is limited and alert burden can undermine use [21]. Counterfactual planning should consequently be evaluated for prescribing errors, omitted harms, delayed treatment, inappropriate continuation or switching, monitoring failures, workload, override patterns, and consequences for patients—not simply for the frequency with which its preferred trajectory is selected.

Responsible clinical decision support further requires transparent intended use, accountable human oversight, monitoring, and governance across the system lifecycle [22]. Technical validation should therefore be followed by causal evaluation, external and temporal validation, silent prospective operation, human-factors testing, medication-safety assessment, equity and transportability analysis, and controlled evaluation of decision and patient outcomes. Progression between stages should remain conditional rather than automatic.

Safe integration therefore requires staged clinical evaluation, patient-important medication-safety assessment, and lifecycle governance rather than model-performance reporting alone [20-22]. No universal readiness score is proposed. A serious failure in causal validity, safety, equity, comprehension, accountability, or monitoring should not be offset by strong performance in another domain.

During use, the output should be presented as a qualified comparison of alternatives, accompanied by assumptions, excluded outcomes, uncertainty, and reasons for any abstention. The interface should separate evidence description from the authorized medication decision. Passing technical evaluation would establish neither clinical utility nor permission to alter treatment; progression would remain conditional on medication-safety, workflow, equity, and governance evidence.

Clinical and Governance Implications

Counterfactual medication planning distributes work across clinical, technical, organizational, and patient roles. Responsible healthcare machine learning requires stakeholder involvement from problem formulation through implementation and monitoring [23]. In this framework, that principle means involving patients and medication professionals when defining outcomes and burdens, not only after a model has been developed.

Pharmacists are positioned to verify medication histories, identify treatment-version ambiguity, assess interactions and adverse-effect pathways, evaluate adherence and access, and design monitoring. Prescribers integrate the comparison with diagnosis, examination, comorbidity management, and the broader therapeutic plan. Patients contribute preferences, lived treatment burden, acceptable uncertainty, and goals that cannot be inferred reliably from structured records. Technical teams remain responsible for data provenance, model behavior, reproducibility, change control, and failure investigation.

Real-world implementation is not equivalent to installing software. Integration studies demonstrate the importance of workflow redesign, role definition, training, stakeholder alignment, and adaptation to local clinical practice [24]. An organization considering the proposed framework would therefore need named clinical, medication-safety, technical, and governance owners; defined escalation routes; documentation standards; and authority to restrict, suspend, or withdraw use.

Trust should be calibrated rather than maximized. Clinician trust in artificial intelligence depends on the user, task, system characteristics, evidence, and organizational context [25]. Excessive trust may produce automation bias, whereas indiscriminate distrust may prevent useful scrutiny of alternatives. Interfaces and training should support appropriate reliance by showing the basis and limitations of each trajectory, not by presenting confidence as certainty or explanation as proof.

The framework proposes that trajectory generation, assumption review, medication-safety interpretation, patient deliberation, treatment authorization, and monitoring be assigned explicitly. These assignments organize responsibility but do not redefine legal liability or professional scope. Organizations remain accountable for determining whether the system's intended use is compatible with local policies, staffing, information systems, patient population, and capacity for continuing oversight.

Limitations

The framework cannot guarantee that explanations correspond to a model's true reasoning or establish causal validity, fairness, safety, or clinical usefulness [26]. Explanations may support scrutiny but can also create false reassurance.

Equity cannot be inferred from aggregate accuracy. Healthcare algorithms may reproduce inequity when proxy outcomes encode unequal access or resource use rather than underlying need [27]. Counterfactual medication planning therefore requires examination of representation, treatment availability, measurement, proxy selection, subgroup performance, and unequal consequences of error.

Transportability is similarly conditional. Clinical and machine-learning findings may not generalize across populations, institutions, formularies, workflows, or treatment practices [28]. Local validation does not remove every limitation, but absence of local evidence restricts interpretation.

Explanation, equity, and transportability are independent decision boundaries, and apparent strength in one domain cannot compensate for failure in another [26-28]. The framework also cannot observe both treatment futures in the same patient, eliminate unmeasured confounding, ensure stable preferences, or represent every rare and delayed consequence. Its architecture and transition rules remain proposed and require empirical evaluation.

Conclusion

Medication planning before treatment is changed requires comparison of plausible futures, not merely prediction of risk under observed care. The proposed Counterfactual Medication-Planning Framework organizes this task around a defined decision time, a verified patient–medication state, feasible treatment strategies, explicit estimands, multidimensional alternative trajectories, an assumption and confounding ledger, an uncertainty envelope, an identifiability and comparability gate, human deliberation, and monitored decision dispositions.

The framework's principal contribution is the separation of processes that are often conflated. Predicting an outcome is not equivalent to estimating a treatment contrast. Constructing a trajectory is not equivalent to identifying its causal effect. Producing an interpretable output is not equivalent to establishing safety or usefulness. Implementing a system is not equivalent to demonstrating patient benefit, and automated analysis does not confer professional authority to change medication.

The framework also treats abstention, restriction, deferment, and information acquisition as legitimate planning outcomes. This is important where data are incomplete, treatment alternatives lack overlap, adherence is uncertain, outcomes are omitted, patient preferences are unstable, or the comparison is materially nonidentifiable. The framework is intended to expose these boundaries rather than conceal them behind a single ranking.

Validation would require evidence across causal methodology, technical performance, medication safety, human factors, patient outcomes, equity, transportability, implementation, and lifecycle governance. Responsibility must remain explicit across pharmacists, prescribers, patients, technical teams, and accountable organizations. Until such validation is completed in defined settings and uses, the framework should be interpreted as an original conceptual architecture for research, evaluation, and disciplined deliberation—not as a clinical recommendation, universal standard, or deployment-ready system.

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

References

1. Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. Nat Med. 2024;30(4):958-68. doi:10.1038/s41591-024-02902-1

2. Smit JM, Krijthe JH, Kant WMR, Labrecque JA, Komorowski M, Gommers DAMPJ, et al. Causal inference using observational intensive care unit data: A scoping review and recommendations for future practice. NPJ Digit Med. 2023;6(1):221. doi:10.1038/s41746-023-00961-1
3. Nguyen TL, Collins GS, Landais P, Le Manach Y. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials—An illustration with the International Stroke Trial. J Clin Epidemiol. 2020;125:47-56. doi:10.1016/j.jclinepi.2020.05.022
4. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell. 2020;2(7):369-75. doi:10.1038/s42256-020-0197-y
5. Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement. Ann Intern Med. 2020;172(1):35-45. doi:10.7326/M18-3667
6. Hoogland J, Efthimiou O, Nguyen TL, Debray TPA. Evaluating individualized treatment effect predictions: A model-based perspective on discrimination and calibration assessment. Stat Med. 2024;43(23):4481-98. doi:10.1002/sim.10186
7. Zhang Y, Laber EB, Davidian M, Tsiatis AA. Interpretable dynamic treatment regimes. J Am Stat Assoc. 2018;113(524):1541-9. doi:10.1080/01621459.2017.1345743
8. Bica I, Alaa AM, Lambert C, van der Schaar M. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. Clin Pharmacol Ther. 2021;109(1):87-100. doi:10.1002/cpt.1907
9. Ruhrberg Estévez S, Peck R, McKinney E, Weatherall J, Bailey S, Rochon J, et al. Causal inference and digital twins: A roadmap for the future of clinical trials. NPJ Digit Med. 2026;9(1):582. doi:10.1038/s41746-026-02871-4
10. Hussain Z, De Brouwer E, Boiarsky R, Setty S, Gupta N, Liu G, et al. Joint AI-driven event prediction and longitudinal modeling in newly diagnosed and relapsed multiple myeloma. NPJ Digit Med. 2024;7(1):200. doi:10.1038/s41746-024-01189-3
11. Allam A, Feuerriegel S, Rebhan M, Krauthammer M. Analyzing patient trajectories with artificial intelligence. J Med Internet Res. 2021;23(12). doi:10.2196/29812
12. Corny J, Rajkumar A, Martin O, Dode X, Lajonchère JP, Billuart O, et al. A machine learning-based clinical decision support system to identify prescriptions with a high risk of medication error. J Am Med Inform Assoc. 2020;27(11):1688-94. doi:10.1093/jamia/ocaa154
13. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. N Engl J Med. 2017;377(14):1391-8. doi:10.1056/NEJMsm1605385
14. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716-20. doi:10.1038/s41591-018-0213-5
15. Gottesman O, Johansson F, Komorowski M, Faisal AA, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25(1):16-8. doi:10.1038/s41591-018-0310-5
16. Li H, Zang C, Xu Z, Pan W, Rajendran S, Chen Y, et al. Federated target trial emulation using distributed observational data for treatment effect estimation. NPJ Digit Med. 2025;8(1):387. doi:10.1038/s41746-025-01803-y
17. Cinelli C, Hazlett C. Making sense of sensitivity: Extending omitted variable bias. J R Stat Soc Series B Stat Methodol. 2020;82(1):39-67. doi:10.1111/rssb.12348
18. Jin Y, Ren Z, Candès EJ. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. Proc Natl Acad Sci U S A. 2023;120(6). doi:10.1073/pnas.2214889120
19. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proc Natl Acad Sci U S A. 2018;115(11):2571-7. doi:10.1073/pnas.1708282114
20. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33. doi:10.1038/s41591-022-01772-9
21. Holbrook AM, Silva JM, Faruque JAY, Deng J, Schneider T, Jaffer A. Effect of electronic drug-drug interaction alerts on patient and clinician outcomes: A systematic review. J Am Med Inform Assoc. 2025;32(10):1617-28. doi:10.1093/jamia/ocaf139
22. Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: Recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc. 2024;31(11):2730-9. doi:10.1093/jamia/ocae209
23. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
24. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. JMIR Med Inform. 2020;8(7). doi:10.2196/15182
25. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: Focus on clinicians. J Med Internet Res. 2020;22(6). doi:10.2196/15154
26. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11). doi:10.1016/S2589-7500(21)00208-9
27. 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):447-53. doi:10.1126/science.aax2342
28. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health. 2020;2(9). doi:10.1016/S2589-7500(20)30186-2