ORIGINAL RESEARCH **Open access**

Deep Reinforcement Learning with Safety Shielding for Personalized Anticoagulation Management in Atrial Fibrillation Patients at High Bleeding Risk Using INR Measurements

Andrei Popescu[1*], Mihai Ionescu[1], Elena Stan[2], Sorin Dumitrescu[1], Irina Pavel[2]

Abstract

Atrial fibrillation affects over 30 million people worldwide and requires long-term anticoagulation, with warfarin still widely used due to its efficacy and reversibility, but its narrow therapeutic window (INR 2.0–3.0) makes dosing particularly challenging, especially in high bleeding-risk patients where both under- and over-anticoagulation can lead to serious complications. Conventional dosing approaches rely on population-based nomograms and clinician judgment, failing to capture individual variability driven by genetics, diet, comorbidities, and drug interactions. To address this limitation, this article proposes a conceptual framework that integrates deep reinforcement learning with a safety-shield mechanism for personalized warfarin dosing. The system uses a deep Q-network trained on historical patient trajectories within an offline Markov Decision Process to recommend dose adjustments based on INR history and clinical risk factors, while a deterministic rule-based safety layer blocks unsafe actions, such as dose increases when INR exceeds 3.5 or extreme adjustments requiring clinician review. Conservative offline reinforcement learning further reduces the risk of unsafe policy extrapolation by limiting overestimation of out-of-distribution actions. Together, this hybrid architecture aims to improve time in therapeutic range while minimizing bleeding risk, providing a structured and clinically constrained approach for safer, individualized anticoagulation management in high-risk atrial fibrillation patients.

Keywords Clinical decision support, Reinforcement learning, Atrial fibrillation, Warfarin dosing, Safety shielding, Offline RL

*Correspondence:
Andrei Popescu
andrei.popescu@gmail.com
[1] Department of Healthcare AI Engineering, University of Bucharest, Bucharest, Romania
[2] Department of Medical Intelligence Systems, Politehnica University of Bucharest, Bucharest, Romania

Introduction

Atrial fibrillation represents the most prevalent sustained cardiac arrhythmia globally, affecting an estimated 33 million individuals and conferring a fivefold increase in thromboembolic stroke risk [1]. Warfarin, a vitamin K antagonist, has constituted the cornerstone of stroke prevention in AFib for over six decades and continues to be widely utilized where direct oral anticoagulants are contraindicated or cost-prohibitive [2]. The clinical challenge centers on a narrow therapeutic window, where INR below 2.0 exposes patients to thromboembolic events while values exceeding 3.0 substantially elevate hemorrhagic risk [3]. Achieving consistent INR control requires individualized dose titration accounting for interpatient variability in drug metabolism, dietary habits, and concomitant medications, a task for which current clinical protocols remain inadequately equipped.

A subset of atrial fibrillation patients faces disproportionately elevated bleeding risk due to clinical factors including advanced age exceeding 75 years, prior intracranial hemorrhage, renal or hepatic dysfunction, concurrent antiplatelet therapy, and labile INR control [4, 5]. Current clinical practice guidelines recognize these patients as requiring intensified monitoring and conservative dosing strategies, yet management protocols lack granularity to dynamically adapt to each patient's evolving risk profile [6]. Standard warfarin nomograms prescribing fixed percentage dose adjustments based on current INR alone treat all patients uniformly regardless of bleeding diathesis and have demonstrated limited efficacy among complex, high-risk individuals [7]. The inadequacy of population-based approaches underscores the need for data-driven personalization simultaneously optimizing therapeutic efficacy and bleeding avoidance.

Reinforcement learning has emerged as a compelling paradigm for sequential clinical decision-making, learning optimal treatment policies through iterative interaction with dynamic patient states and delayed outcome feedback [8, 9]. In warfarin management, RL agents can model the complex temporal relationship between dose adjustments and subsequent INR responses, potentially identifying personalized dosing strategies outperforming static protocols. However, deployment in high-stakes environments raises critical safety concerns, as standard algorithms may explore hazardous dose adjustments during learning [10, 11]. The exploration-exploitation tradeoff becomes ethically untenable when suboptimal actions could precipitate intracranial hemorrhage or ischemic stroke [12].

This article proposes a conceptual framework addressing the dual imperatives of personalization and safety through integration of deep reinforcement learning with clinician-defined safety shielding for warfarin dose optimization in high-risk atrial fibrillation patients [13]. The framework operates exclusively offline, learning treatment policies from retrospective INR trajectories without real-time patient interaction. By embedding a deterministic safety layer enforcing hard clinical constraints on RL policy outputs, the architecture provides verifiable safety guarantees while enabling personalized dose selection informed by each patient's unique response pattern. The subsequent sections detail clinical background, formal problem formulation, architectural components, and evaluation strategy for this safe RL approach.

Background

Warfarin pharmacokinetics and INR monitoring

Warfarin exerts its anticoagulant effect through inhibition of vitamin K epoxide reductase complex subunit 1, disrupting cyclic interconversion of vitamin K and impairing hepatic synthesis of functional coagulation factors II, VII, IX, and X [14]. The drug exhibits near-complete oral absorption, extensive protein binding exceeding 99%, and hepatic metabolism primarily via cytochrome P450 2C9, with genetic polymorphisms accounting for approximately 35–50% of interindividual dose variability [15]. Therapeutic monitoring relies on the INR, a standardized measure of the extrinsic coagulation pathway. The established therapeutic range for atrial fibrillation is INR 2.0–3.0, with values below 2.0 indicating insufficient anticoagulation and heightened stroke vulnerability, while INR exceeding 4.0 confers exponentially increasing risk of major hemorrhage, particularly intracranial bleeding [16]. The temporal relationship between dose adjustment and INR effect exhibits a 36–72 hour delay, complicating real-time titration and necessitating anticipatory strategies informed by both current and prior INR trends [17].

High bleeding risk definition and clinical implications

High bleeding risk in atrial fibrillation patients is systematically assessed through validated instruments, most notably the HAS-BLED score incorporating hypertension, abnormal renal or hepatic function, prior stroke, previous major bleeding, labile INR, age greater than 65 years, and concomitant antiplatelet or NSAID use [18]. A HAS-BLED score of three or greater identifies patients at substantially elevated hemorrhage risk, with prior intracranial hemorrhage representing a particularly strong predictor of recurrent bleeding events [19]. Current consensus guidelines from European and Asia-Pacific expert panels emphasize that high bleeding risk does not constitute an absolute anticoagulation contraindication but mandates intensified monitoring, modifiable risk factor mitigation, and careful agent selection [20]. For these patients, the safety margin for dosing errors narrows considerably, as supratherapeutic INR excursions more readily translate into clinically significant bleeding, rendering algorithmic recommendations without explicit safety constraints potentially hazardous. Current bleeding risk prediction tools, including the DOAC Score for patients

on direct oral anticoagulants, highlight the continued need for individualized risk assessment approaches [21].

Current warfarin dosing protocols and their limitations

Contemporary warfarin dosing protocols typically employ nomogram-based algorithms prescribing fixed-percentage dose adjustments based on current INR relative to target range, with typical adjustments ranging from 10% to 20% depending on the degree of deviation [22]. The initial dosing phase utilizes loading strategies of 5–10 mg daily, followed by maintenance dosing guided by serial INR measurements obtained every one to four weeks [23]. Machine learning models have been developed to predict stable warfarin maintenance doses using clinical and pharmacogenetic features, including deep neural networks integrating CYP2C9 and VKORC1 genotype data with demographic variables [24, 25]. Multivariate linear regression and machine learning algorithms have been compared for precision warfarin dosing prediction in various populations, demonstrating that algorithmic approaches can improve dosing accuracy over standard clinical methods [26]. However, these static prediction approaches fail to dynamically adapt to evolving temporal patterns of individual INR response, and nomogram protocols were predominantly validated in general AFib populations with limited data in high bleeding risk cohorts where dosing error consequences are most severe.

Offline reinforcement learning for clinical decision support

Offline reinforcement learning, also termed batch RL, addresses limitations of standard online RL by learning treatment policies exclusively from previously collected observational datasets without live clinical environment interaction [27]. This paradigm suits healthcare applications where random exploration of dangerous actions during learning is ethically and practically infeasible. Conservative Q-learning modifies the standard Q-function objective by adding a penalty discouraging overestimation of action values for state-action pairs poorly represented in training data, thereby producing policies within observed clinical practice support [28]. In critical care, offline RL has demonstrated capacity to learn treatment strategies suggesting survival benefits over clinician decision-making when evaluated retrospectively [29, 30]. The methodological framework established for offline RL in healthcare provides rigorous foundation for extending these approaches to warfarin dosing, where abundant historical INR trajectories exist in anticoagulation clinic databases. Guidelines for reinforcement learning in healthcare have emphasized the importance of safety considerations, interpretability, and rigorous evaluation prior to clinical deployment [31].

Framework Overview

High-level architecture

The proposed framework comprises a sequential decision pipeline wherein the patient's anticoagulation state—encompassing current INR, recent warfarin doses, and bleeding risk factors—is first processed by a deep Q-network that evaluates candidate dose adjustment actions and selects the option maximizing expected long-term therapeutic benefit [1]. This initial dose recommendation passes through a safety shielding module applying deterministic clinical rules to mask any actions considered hazardous given current INR and patient risk profile [2]. The shield operates as a hard constraint layer, deterministically eliminating dose increase actions when INR exceeds prespecified supratherapeutic thresholds and completely blocking all automated recommendations to mandate clinician review when INR reaches critical values [3]. This gated architecture ensures personalization capabilities are preserved for safe operating regions while guaranteeing no unsafe action can reach clinical implementation, establishing a least-privilege principle for algorithmic recommendations. The complete system is designed for offline deployment, processing batch historical data to generate treatment policies subsequently applicable to new patient encounters within the training data distribution.

Figure 1 presents the proposed safety-shielded offline deep reinforcement learning architecture for personalized warfarin dose management in high-bleeding-risk atrial fibrillation patients.

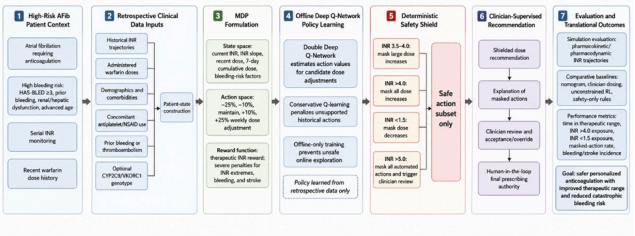

Figure 1. Safety-Shielded Offline Deep Reinforcement Learning Framework for Personalized Warfarin Management in High-Bleeding-Risk Atrial Fibrillation Patients.

Core assumptions

The framework operates under several foundational assumptions defining its applicability scope and guiding design [4]. First, it assumes availability of a historical dataset containing sequential INR measurements, administered warfarin doses, and clinical outcomes from atrial fibrillation patients managed under standard care, with adequate representation of high bleeding risk individuals. Second, clinically defined safety constraints derived from established anticoagulation guidelines are available in formalized, computationally tractable representation amenable to integration as action masking rules [5]. Third, the framework is designed exclusively for offline training and batch evaluation, with no provision for online policy updates during deployment without separate regulatory approval, prospective validation, and institutional review [6]. Fourth, the learned policy serves as a clinical decision support tool rather than an autonomous dosing agent, with final dose selection remaining the treating clinician's responsibility. Systematic reviews of machine learning for warfarin therapy have highlighted the importance of these assumptions in ensuring safe and effective clinical translation of algorithmic approaches [7].

Design principles

Four overarching principles govern the framework's architecture [8]. The safety-first principle establishes that preventing harm supersedes optimizing efficacy, operationalized through the safety shield's unconditional authority to override RL outputs and through asymmetric reward penalties heavily penalizing bleeding-associated actions. The personalization principle requires dosing policies to adapt to individual patient characteristics, including genetic variants when available, temporal INR response patterns, and bleeding risk factors modulating acceptable risk-benefit calculus. The offline-only principle mandates all policy learning occurs on retrospective data, with conservative Q-learning ensuring no extrapolation beyond observed clinical strategies [9]. The interpretability principle demands transparency in safety constraints applied, such that clinicians understand precisely which rule triggered an action rejection and why a recommendation was modified. Reinforcement learning surveys in healthcare have emphasized that these principles are essential for building clinician trust and enabling appropriate oversight of algorithmic recommendations in high-stakes clinical environments [10].

Markov Decision Process Formulation

State space definition

The state representation encodes information available to clinicians at each dosing decision point, capturing both anticoagulation status and bleeding risk [11]. Primary state features include current INR measurement, most recent warfarin dose in milligrams, cumulative dose over the preceding seven days, and INR trend operationalized as the difference and slope between current and prior INR values obtained at intervals of three to seven days. Additional clinical features comprise HAS-BLED score components: uncontrolled hypertension, abnormal renal function indicated by serum creatinine exceeding 200 μmol/L or chronic dialysis, abnormal hepatic function denoted by cirrhosis or transaminases exceeding twice normal, patient age, and concurrent antiplatelet or NSAID use [12]. Prior major bleeding events including intracranial hemorrhage and gastrointestinal bleeding requiring transfusion are encoded as binary flags. When available, CYP2C9 and VKORC1 genotype data can be included as categorical variables, though the framework functions with or without pharmacogenetic information. Machine learning models for warfarin dose prediction have demonstrated that incorporating these clinical features significantly improves dosing accuracy across diverse populations [13].

Action space design

The action space defines warfarin dose adjustments available to the RL policy at each decision point, structured to reflect clinical standard titration granularity [14]. The primary action space comprises five discrete dose adjustment categories: 25% dose decrease, 10% dose decrease, maintenance of current dose, 10% dose increase, and 25% dose increase applied to the patient's current total weekly warfarin dose. This discretization reflects typical clinical practice while maintaining tractable action space for Q-learning algorithms. The framework generalizes to continuous action representation where the policy outputs a continuous scaling factor, though discrete formulation aligns with modular safety shielding and

facilitates direct comparison with standard clinical protocols [15]. The action space is defined prior to safety shield application, meaning the full set remains available during Q-value computation, with unsafe actions subsequently masked; this preserves the ability to learn Q-values for safe actions becoming preferable when dangerous alternatives are eliminated. Studies optimizing warfarin dosing using deep reinforcement learning have employed similar action space discretizations, demonstrating their clinical validity for dose adjustment modeling [16].

Reward function definition

The reward function mathematically encodes clinical objectives of warfarin therapy, assigning numerical values reflecting outcome desirability [17]. The primary structure assigns +1 when subsequent INR falls within therapeutic range 2.0–3.0, a modest penalty of −0.5 for near-therapeutic values in ranges 1.5–2.0 or 3.0–4.0 representing suboptimal but not immediately dangerous anticoagulation, and a severe penalty of −2.0 for critically subtherapeutic INR below 1.5 or markedly supratherapeutic INR exceeding 4.0. A catastrophic penalty of −5.0 is assigned for major bleeding events including intracranial hemorrhage, retroperitoneal bleeding, or bleeding requiring hospitalization, while ischemic stroke or systemic embolism incurs symmetric −5.0 penalty. This asymmetric structure heavily penalizes the most dangerous outcomes while providing gradient signal for intermediate INR deviations [18]. Optimization frameworks for warfarin management have emphasized that reward function design must reflect the asymmetric clinical consequences of bleeding versus thromboembolic events, particularly for high-risk patients where hemorrhage represents the dominant clinical concern [19].

Deep Reinforcement Learning Architecture

Q-network design

The Q-network architecture employs a feedforward deep neural network comprising an input layer ingesting the concatenated patient state vector, two to three fully connected hidden layers of 128 to 256 units each with rectified linear unit activation functions, and an output layer producing a scalar Q-value for each discrete action in the dose adjustment space [20]. The network is trained using double deep Q-learning to mitigate the overestimation bias inherent to standard Q-learning, where action selection and evaluation are decoupled by using the online network to select the maximizing action and the target network to evaluate its value. The target network parameters are periodically updated via soft copying from the online network with a Polyak averaging coefficient of 0.005, promoting training stability in the face of non-stationary target values characteristic of bootstrapped temporal difference learning [21]. The loss function minimizes mean squared error between predicted Q-value and temporal difference target, with an additional conservative penalty term specific to the offline setting. Deep neural network architectures for warfarin dose decision support have demonstrated capacity to capture complex nonlinear relationships between patient features and optimal dosing, providing foundation for RL-based extensions [22].

Policy selection mechanism

During training, action selection follows an epsilon-greedy strategy with initial exploration rate of 0.3 decaying exponentially to a terminal value of 0.05 over the course of training, ensuring sufficient state-action space exploration while asymptotically favoring exploitation of learned knowledge [23]. At evaluation and deployment time, the policy employs greedy action selection, deterministically choosing the dose adjustment action with maximum Q-value from the set permitted by the safety shield. The conservative Q-learning objective modifies standard Q-learning by adding a penalty minimizing Q-values for actions outside the behavioral policy's support, computed as the difference between Q-value of the selected action and expected Q-value under the data-generating policy, preventing preference for actions with insufficient training data [24]. This conservative mechanism is essential in warfarin dosing where certain dose adjustments may be rarely observed in historical data, particularly large dose increases in high-risk patients, and where overestimation could lead to dangerous recommendations if deployed without the safety shield. Reinforcement learning approaches for optimizing dynamic treatment regimes have demonstrated that conservative policy selection is critical for maintaining safety when transitioning from offline training to clinical evaluation [25].

Safety Shielding

Action masking rules

The safety shield implements a set of deterministic, clinician-defined action masking rules that operate on current INR values and patient bleeding risk status to prevent hazardous dose recommendations from reaching clinical implementation [1]. When INR exceeds 4.0, all dose increase actions are unconditionally masked, restricting the policy to maintenance or decrease options only, reflecting the clinical imperative to reduce anticoagulation intensity in the setting of supratherapeutic INR. For INR values in the 3.5–4.0 range, large increase actions of 25% are masked while conservative 10% decreases and maintenance remain available, acknowledging that mild supratherapeutic excursions may warrant only modest dose reduction rather than aggressive de-escalation. When INR falls below 1.5, all dose decrease actions are masked, preventing the policy from further reducing anticoagulation in patients already at heightened thromboembolic risk due to inadequate anticoagulation. Critically, when INR exceeds 5.0, all automated actions are masked and a physician alert is triggered, removing algorithmic dosing authority entirely and mandating immediate clinical assessment consistent with guideline-directed management of severe over-anticoagulation [2]. These threshold-based rules provide a transparent, auditable safety layer whose behavior is fully specified and verifiable independent of the underlying Q-network's training state.

Table 1 formalizes the deterministic safety shield as a clinically interpretable action-masking layer that constrains the reinforcement learning policy before any dose recommendation reaches the clinician.

Table 1. Clinical Safety Shield Logic for State-Dependent Warfarin Dose Action Masking

Clinical state condition	Permitted automated action space	Masked action(s)	Safety ratio
INR within therapeutic range, 2.0–3.0	−10%, maintain, +10%; larger changes only if clinically justified by state history	No automatic masking unless patient-specific risk rule applies	Therapeu range perr personaliza while avoi unnecess aggressive change
INR 3.5–4.0	Maintain, −10%, −25%	+25% increase	Mild-to-mod supratherap anticoagula
			narrows t safety marg high-risk pa
INR >4.0	Maintain, −10%, −25%	+10%, +25% increases	Supratherap INR substar increase hemorrhagi
INR >5.0	No automated dose recommendation	All dose actions	Severe ov anticoagula requires di clinical assessme rather tha algorithm dosing
INR <1.5	Maintain, +10%, +25%	−10%, −25% decreases	Critically subtherape INR increa thromboem risk
Prior intracranial hemorrhage or extreme HAS-BLED profile	Conservative subset only, institution-defined	Aggressive increase actions unless explicitly cleared	Catastrop bleeding his requires str tolerance supratherap exposur
Out-of-distribution state detected by offline model	Rule-based or clinician-review pathway	RL-selected action if unsupported by historical data	Offline RL overestim poorly represent actions

Constrained policy optimization

The safety shield formalizes constrained policy optimization by restricting the RL policy's action selection to a state-dependent safe action subset defined by clinical rules, effectively projecting the unconstrained policy onto the feasible action space at each decision point [3]. This approach represents a hard constraint formulation distinct from Lagrangian relaxation methods that encode safety as soft penalties in the reward function, as hard constraints provide categorical safety guarantees that soft constraints cannot ensure under distribution shift or model misspecification. The constrained optimization can be expressed as the standard RL objective subject to the

constraint that for all states, the probability of selecting an unsafe action is identically zero, a condition enforced by the shield's deterministic masking operation. Prior theoretical work on safe reinforcement learning through shielding has established that such hard constraint approaches provide formal safety guarantees, ensuring that the composite system cannot violate prespecified safety properties regardless of the underlying policy network's behavior [4]. In the warfarin dosing context, this formalism translates to an absolute guarantee that supratherapeutic INR values will never trigger dose increase recommendations, directly preventing the most dangerous class of algorithmic dosing errors.

Clinical rule integration

The safety shielding rules are derived from consensus clinical guidelines for warfarin management, incorporating threshold-based recommendations from the American College of Chest Physicians and European Society of Cardiology guidelines that specify appropriate dose adjustments for various INR ranges [5, 6]. Clinician-defined thresholds govern shield behavior, allowing institutional customization of safety boundaries to reflect local practice patterns, patient population characteristics, and anticoagulation clinic protocols. The shield's decision logic is inherently explainable, as each action rejection is accompanied by a specific clinical rule citation that can be surfaced to the treating clinician, such as "dose increase masked: INR 4.2 exceeds threshold for dose escalation; current guidelines recommend dose reduction per protocol." This transparency stands in contrast to the opaque decision-making of the neural network policy and supports appropriate trust calibration, enabling clinicians to understand precisely when and why the algorithm's recommendation was modified by safety constraints. The principle of constrained policy optimization within clinical decision support has been advanced as essential for translating RL advances into safe bedside tools, with the warfarin domain exemplifying how domain-specific clinical knowledge can be formalized as verifiable computational constraints [7].

Offline Training

Data requirements

The framework's offline training procedure requires a comprehensive historical dataset comprising sequential INR measurements, administered warfarin doses, demographic and clinical covariates, and clinical outcomes from patients with atrial fibrillation managed in anticoagulation clinics [8]. A dataset containing between 10,000 and 100,000 patient-timepoints—representing approximately 1,000 to 5,000 distinct patient episodes with multiple dosing encounters per patient—provides sufficient sample size for deep Q-network training while capturing the heterogeneity of INR response patterns across diverse patient populations. Each data record must include the current INR, the warfarin dose preceding that INR measurement, the dose prescribed in response, the subsequent INR, and clinical covariates including age, renal and hepatic function parameters, bleeding history, and concomitant medications known to interact with warfarin metabolism. Datasets utilized in prior warfarin machine learning studies, including those derived from the MIMIC-III critical care database and institutional anticoagulation registries, demonstrate the feasibility of constructing such training corpora [9, 10]. The retrospective nature of these data imposes limitations regarding unmeasured confounders and selection bias, which must be addressed through careful study design and acknowledged in interpreting learned policies.

Conservative Q-learning implementation

The training objective employs Conservative Q-Learning, which augments the standard temporal difference loss with a regularization term that minimizes Q-values for actions outside the support of the behavioral policy that generated the historical data [11]. This penalty is computed as the difference between the Q-value of the action selected by the learned policy and the expected Q-value under the empirical behavioral distribution observed in the dataset, thereby penalizing the learned policy for favoring actions rarely or never taken by historical clinicians. The conservative penalty strength, governed by a hyperparameter alpha controlling the bias-variance tradeoff between conservatism and optimism, must be tuned on held-out validation data to balance safety against potential under-treatment. In the context of high-risk warfarin patients, conservative regularization is particularly important for dose increase actions, as clinicians may have rarely escalated doses in patients perceived as high bleeding risk, creating a systematic underrepresentation of aggressive dosing strategies in observational data that could otherwise be erroneously undervalued by a naive offline RL algorithm. Offline reinforcement learning methods

incorporating conservatism principles have demonstrated robust performance in clinical decision support settings where exploration is impossible and policy value must be estimated from fixed datasets [12].

Evaluation Strategy

Simulation environment

Evaluation of the trained safety-shielded RL policy proceeds through a warfarin dose response simulator implementing a pharmacokinetic/pharmacodynamic model that generates synthetic INR trajectories in response to administered warfarin doses [13]. The simulator incorporates population-level warfarin pharmacokinetic parameters including absorption rate, volume of distribution, and hepatic clearance, along with a pharmacodynamic model linking warfarin concentration to INR response through inhibition of vitamin K-dependent clotting factor synthesis with appropriate half-lives for factors II, VII, IX, and X [14]. Patient-specific parameters are sampled from distributions derived from published population pharmacokinetic studies, enabling simulation of heterogeneous virtual patients spanning the spectrum of warfarin sensitivity observed clinically, including those with CYP2C9 and VKORC1 variants conferring altered dose requirements. Bleeding and thromboembolic risk models are integrated into the simulator, generating probabilistic adverse events as functions of cumulative time in supratherapeutic and subtherapeutic INR ranges, respectively, calibrated to event rates reported in major atrial fibrillation clinical trials [15]. The simulator enables head-to-head comparison of the safety-shielded RL policy against standard clinical protocols across thousands of simulated patient episodes without exposing real patients to algorithmic dosing, serving as the primary evaluation platform prior to retrospective clinical data analysis.

Performance metrics

The primary performance metric is the percentage time in therapeutic range, defined as the proportion of simulated INR measurements falling within the 2.0–3.0 range, recognized as the standard quality indicator for anticoagulation control and validated as a surrogate endpoint correlating with clinical outcomes [16]. Secondary safety metrics include the percentage of time with INR above 4.0, which directly quantifies exposure to the supratherapeutic range most strongly associated with hemorrhagic complications, and the percentage of time with INR below 1.5, capturing periods of critically inadequate anticoagulation conferring thromboembolic vulnerability. The safety shield's operational impact is quantified through the percentage of RL policy actions that are masked by the shield, reflecting the frequency with which the unconstrained policy would have recommended potentially dangerous dose adjustments and providing a direct measure of the shield's clinical relevance. Additionally, the cumulative incidence of simulated major bleeding events and ischemic strokes over extended virtual follow-up periods provides clinical outcome measures that integrate the temporal dynamics of anticoagulation control into interpretable patient-centered endpoints [17].

Baseline comparisons

The safety-shielded RL policy is benchmarked against three comparators representing the spectrum of current and potential future anticoagulation management strategies [18]. The standard warfarin nomogram comparator implements a guideline-based dosing algorithm that prescribes fixed-percentage dose adjustments based solely on the current INR value, representing the status quo in most anticoagulation clinics. A clinician dosing comparator replicates the empirical dosing behavior observed in the historical dataset, reflecting real-world clinical practice with its inherent variability and experiential judgment. The unconstrained RL comparator evaluates the deep Q-network policy without the safety shield, quantifying the safety benefit attributable specifically to the shielding mechanism and assessing whether improved therapeutic range performance, if any, comes at the cost of increased unsafe action selection. A rule-based safety-only comparator applies the shielding rules without any RL personalization, effectively serving as a conservative clinical protocol that never recommends unsafe actions but also lacks the capacity to learn individualized dosing patterns from patient response history, thereby isolating the personalization benefit of the RL component [19].

Table 2 clarifies how the proposed framework distributes responsibility across personalization, conservatism, deterministic safety enforcement, explainability, and clinician oversight.

Table 2. Analytical Separation of Personalization, Safety, Conservatism, and Clinical Oversight in the Proposed Framework

Framework layer	Primary function	Mechanism of control	Clinic ac
MDP state representation	Converts anticoagulation management into sequential decision states	Encodes INR history, dose history, INR trend, bleeding-risk factors, and optional genotype	Cap patien ten respon than pati homo
Deep Q-network	Learns individualized dose-adjustment preferences	Estimates long-term value of discrete dose actions	En perso titratic on pr res pa
Conservative offline RL	Restricts learning to historically supported clinical behavior	Penalizes unsupported or rarely observed state-action pairs	Reduc extra f retros c
Deterministic safety shield	Blocks clinically unsafe actions regardless of Q-network output	Applies hard INR- and risk-based action masks	Pro audital guar be recomr rel
Rule explanation layer	Makes safety interventions transparent	Reports which shield rule modified or blocked the RL action	Sup clinic a accou
Human-in-the-loop review	Preserves clinician authority over final prescribing	Allows acceptance, modification, or override of shielded recommendation	Mai me respo and c jud
Simulation and retrospective evaluation	Tests policy behavior before prospective use	Compares against nomograms, clinician behavior, unconstrained RL, and safety-only rules	Sep perso bene safet be

Limitations

Technical limitations

The framework depends on historical dataset quality; systematic biases, missing INR measurements, and underrepresentation of patient subgroups may propagate into the learned policy. The safety shield may prove overly conservative, blocking appropriate dose escalations in patients with rapid warfarin metabolism. The pharmacometric simulator simplifies real-world biology, omitting factors like acute illness, dietary fluctuations, and medication adherence that materially influence INR trajectories. Rigorous sensitivity analyses are needed to assess model robustness under varying data quality assumptions.

Clinical limitations

Unmeasured factors such as dietary vitamin K intake, alcohol consumption, intercurrent illness, and supplement use significantly influence warfarin requirements but remain incompletely captured in electronic health records. Prospective clinical validation through randomized trials is essential before deployment, though resource-intensive and potentially lengthy. Liability concerns surrounding adverse events following algorithmic recommendations represent a significant adoption barrier under current regulatory frameworks. A human-in-the-loop paradigm with transparent documentation of algorithmic recommendation acceptance or override represents the most prudent deployment strategy pending substantial safety evidence.

Conclusion

This article presented a conceptual framework integrating deep reinforcement learning with safety shielding for personalized warfarin management in high bleeding risk atrial fibrillation patients. The framework resolves the tension between machine learning personalization and clinical safety by embedding a deterministic shielding layer guaranteeing adherence to established dosing constraints regardless of neural network outputs.

The key advantages center on a safety-first architecture that categorically prevents dose escalations during

supratherapeutic INR states through transparent action masking rules. Offline training eliminates ethical and logistical barriers to clinical RL deployment, extracting policies from historical anticoagulation data. Personalization via deep Q-learning enables discovery of individualized dosing patterns, potentially achieving superior therapeutic range performance over static nomograms.

Important limitations include dependence on high-quality training data, simulator fidelity constraints, and unmeasured clinical confounders. The safety shield may prove overly restrictive for specific subgroups and requires explicit clinician-defined thresholds balancing protection against appropriate therapeutic aggression. These limitations motivate a phased approach prioritizing simulation evaluation followed by retrospective validation before prospective clinical investigation.

The pathway toward implementation requires collaboration across machine learning, clinical pharmacology, and health system leadership to assemble data infrastructure, validate simulators, and design governance frameworks. Large integrated health systems and anticoagulation clinics with mature electronic health records represent promising settings for initial retrospective evaluation. The vision of safe, personalized anticoagulation through reinforcement learning remains aspirational but grounded in methodological advances making it incrementally achievable.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 15 Dec 2024 Revised: 24 Jan 2025 Accepted: 04 Mar 2025
Published online: 20 July 2025

References

Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomström-Lundqvist C, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J. 2021;42(5):373-498.

Armbruster AL, Benjamin EJ, Chyou JY, Goldberger ZD, Gopinathannair R, Gorenek B, et al. 2023 ACC/AHA/ACCP/HRS guideline for the diagnosis and management of atrial fibrillation: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2023;148(1):e00-e00.

Petch J, Nelson W, Wu M, Ghassemi M, Benz A, Fatemi M, et al. Optimizing warfarin dosing for patients with atrial fibrillation using machine learning. Sci Rep. 2024;14(1):4516.

Maciorowska M, Uziębło-Życzkowska B, Gorczyca-Głowacka I, Wożakowska-Kapłon B, Jelonek O, Wójcik M, et al. Oral anticoagulation therapy in atrial fibrillation patients at high risk of bleeding: clinical characteristics and treatment strategies based on data from the Polish multicenter register of atrial fibrillation (POL-AF). Pol Heart J (Kardiol Pol). 2024;82(1):37-45.

Gorog DA, Gue YX, Chao TF, Fauchier L, Ferreiro JL, Huber K, et al. Assessment and mitigation of bleeding risk in atrial fibrillation and venous thromboembolism: executive summary of a European and Asia-Pacific expert consensus paper. Thromb Haemost. 2022;122(10):1625-52.

Aggarwal R, Ruff CT, Virdone S, Perreault S, Kakkar AK, Palazzolo MG, et al. Development and validation of the DOAC score: a novel bleeding risk prediction tool for patients with atrial fibrillation on direct-acting oral anticoagulants. Circulation. 2023;148(12):936-46.

Ravvaz K, Weissert JA, Ruff CT, Chi CL, Tonellato PJ. Personalized anticoagulation: optimizing warfarin management using genetics and simulated clinical trials. Circ Cardiovasc Genet. 2017;10(6):e001804.

Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Comput Surv. 2021;55(1):1-36.

Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement learning for clinical decision support in critical care: comprehensive review. J Med Internet Res. 2020;22(7):e18477.

Coronato A, Naeem M, De Pietro G, Paragliola G. Reinforcement learning for intelligent healthcare applications: a survey. Artif Intell Med. 2020;109:101964.

Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25(1):16-8.

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.

Zadeh SA, Street WN, Thomas BW. Optimizing warfarin dosing using deep reinforcement learning. J Biomed Inform. 2023;137:104267.

Li X, Li D, Wu JC, Liu ZQ, Zhou HH, Yin JY. Precision dosing of warfarin: open questions and strategies. Pharmacogenomics J. 2019;19(3):219-29.

Roche-Lima A, Roman-Santiago A, Feliu-Maldonado R, Rodriguez-Maldonado J, Nieves-Rodriguez BG, Carrasquillo-Carrion K, et al. Machine learning algorithm for predicting warfarin dose in Caribbean Hispanics using pharmacogenetic data. Front Pharmacol. 2020;10:1550.

Steiner HE, Giles JB, Patterson HK, Feng J, El Rouby N, Claudio K, et al. Machine learning for prediction of stable warfarin dose in US Latinos and Latin Americans. Front Pharmacol. 2021;12:749786.

Lee H, Kim HJ, Chang HW, Kim DJ, Mo J, Kim JE, et al. Development of a system to support warfarin dose decisions using deep neural networks. Sci Rep. 2021;11(1):14745.

Nguyen HD, Cho YS, Kim HS, Han IY, Kim DK, Ahn S, et al. Comparison of multivariate linear regression and a machine learning algorithm developed for prediction of precision warfarin dosing in a Korean population. J Thromb Haemost. 2021;19(7):1676-86.

Choi H, Kang HJ, Ahn I, Gwon H, Kim Y, Seo H, et al. Machine learning models to predict the warfarin discharge dosage using clinical information of inpatients from South Korea. Sci Rep. 2023;13(1):22461.

Fülöp P, Tóth Š, Porubän T, Fülöpová Z, Borovská A, Dvorožňáková M. Machine learning for warfarin therapy: a systematic review. Pharmaceuticals. 2025;18(10):1544.

Roggeveen L, El Hassouni A, Ahrendt J, Guo T, Fleuren L, Thoral P, et al. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. Artif Intell Med. 2021;112:102003.

Wani AA, Abeer F. Application of machine learning techniques for warfarin dosage prediction: a case study on the MIMIC-III dataset. PeerJ Comput Sci. 2025;11:e2612.

Tu R, Luo Z, Pan C, Wang Z, Su J, Zhang Y, et al. Offline safe reinforcement learning for sepsis treatment: tackling variable-length episodes with sparse rewards. Hum Centric Intell Syst. 2025;5(1):63-76.

Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In: Machine Learning for Healthcare Conference. PMLR; 2017. p. 147-63.

Zeng J, Shao J, Lin S, Zhang H, Su X, Lian X, et al. Optimizing the dynamic treatment regime of in-hospital warfarin anticoagulation in patients after surgical valve replacement using reinforcement learning. J Am Med Inform Assoc. 2022;29(10):1722-32.

Ji H, Gill M, Draper EW, Liedl DA, Hodge DO, Houghton DE, et al. Warfarin dose management using offline deep reinforcement learning. IEEE J Biomed Health Inform. 2025 Feb 24. [Epub ahead of print].

Wang Y, Liu A, Yang J, Wang L, Xiong N, Cheng Y, et al. Clinical knowledge-guided deep reinforcement learning for

sepsis antibiotic dosing recommendations. Artif Intell Med. 2024;150:102811.

Achiam J, Held D, Tamar A, Abbeel P. Constrained policy optimization. In: International Conference on Machine Learning. PMLR; 2017. p. 22-31.

Alshiekh M, Bloem R, Ehlers R, Könighofer B, Niekum S, Topcu U. Safe reinforcement learning via shielding. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1).

Asiimwe IG, Zhang EJ, Osanlou R, Jorgensen AL, Pirmohamed M. Warfarin dosing algorithms: a systematic review. Br J Clin Pharmacol. 2021;87(4):1717-29.

Dhippayom T, Boonpattharatthiti K, Kategeaw W, Hong H, Chaiyakunapruk N, Barnes GD, et al. Comparative effectiveness of warfarin management strategies: a systematic review and network meta-analysis. EClinicalMedicine. 2024;74:102614.