hemosim: Gymnasium Environments for Reinforcement Learning in Hemostasis and Anticoagulation Management

Hass Dhia
Smart Technology Investments Research Institute
hass@smarttechinvest.com

April 2026

Abstract

Anticoagulation therapy affects millions of patients worldwide, yet dosing errors contribute to approximately 33,000 emergency department visits annually in the United States alone. Despite the inherently sequential nature of anticoagulant dose titration – where each decision depends on evolving patient state – no standardized reinforcement learning (RL) simulation environments exist for hemostasis and anticoagulation management. We introduce `hemosim`, an open-source Python package providing four Gymnasium-compatible environments built on mechanistic pharmacokinetic/pharmacodynamic (PK/PD) models: warfarin dose titration with CYP2C9/VKORC1 pharmacogenomics, unfractionated heparin infusion management, direct oral anticoagulant (DOAC) selection and dosing for atrial fibrillation, and disseminated intravascular coagulation (DIC) management with multi-component blood product therapy. Proximal Policy Optimization (PPO) agents trained on these environments outperform guideline-based clinical baselines across all four tasks, with the largest improvement observed in DOAC management where PPO achieves a mean reward of 23.80 compared to 12.98 for the clinical baseline (83.4% improvement). `hemosim` is available under the MIT license (`pip install hemosim`), includes 142 unit and integration tests, and provides a standardized benchmark for reproducible research in RL-driven anticoagulation.

1 Introduction

Anticoagulation therapy is among the most common and consequential pharmacological interventions in modern medicine. Warfarin remains the most widely prescribed oral anticoagulant globally despite the emergence of direct oral anticoagulants (DOACs), and unfractionated heparin is the standard parenteral anticoagulant for acute thrombotic events [Hirsh et al., 2001]. Together, anticoagulants are prescribed to tens of millions of patients for conditions including atrial fibrillation, venous thromboembolism, mechanical heart valves, and disseminated intravascular coagulation (DIC).

The clinical challenge of anticoagulation lies in its narrow therapeutic window. Insufficient anticoagulation permits thromboembolic events – stroke, pulmonary embolism, deep vein thrombosis – while excessive anticoagulation causes hemorrhage, which can be fatal. Warfarin's therapeutic index is notoriously narrow: the international normalized ratio (INR) target range of 2.0–3.0 leaves little margin for error, and patient response varies dramatically based on genetics (CYP2C9, VKORC1), diet, concomitant medications, and acute illness [International Warfarin Pharmacogenetics Consortium, 2009, Johnson et al., 2017]. Heparin dosing faces analogous challenges with activated partial

thromboplastin time (aPTT) targets of 60–100 seconds [Raschke et al., 1993]. DOACs, while requiring less routine monitoring, demand drug and dose selection based on renal function, stroke risk, and bleeding risk [Connolly et al., 2009, Patel et al., 2011, Granger et al., 2011]. DIC represents the most complex scenario: simultaneous hemorrhage and thrombosis requiring coordinated multi-component therapy [Levi and Ten Cate, 1999, Levi et al., 2009].

These characteristics make anticoagulation management a natural candidate for reinforcement learning (RL). The problem is inherently sequential: a clinician observes the patient state (laboratory values, clinical parameters), selects an action (dose, drug), and the patient transitions to a new state determined by pharmacokinetics and pharmacodynamics. The delayed and stochastic nature of drug effects further aligns with the RL formulation of sequential decision-making under uncertainty.

Prior work has demonstrated the potential of RL for clinical dosing. Nemati et al. [2016] applied deep RL to heparin dosing from electronic health records, and Anzabi Zadeh et al. [2023] surveyed RL approaches for warfarin dosing. Tosca et al. [2024] reviewed RL applications across clinical pharmacology more broadly. However, a critical limitation pervades this literature: each study constructs ad-hoc simulation environments that cannot be reproduced or compared across research groups. The lack of standardized benchmarks impedes scientific progress in the same way that the absence of Atari or MuJoCo environments would have hindered general RL research.

`hemosim` addresses this gap. We contribute:

1. Four Gymnasium-compatible environments spanning the major anticoagulation modalities, each grounded in published PK/PD models.
2. Mechanistic patient generation with pharmacogenomic variability, enabling biologically plausible episode diversity.
3. Clinically motivated reward functions that balance therapeutic efficacy, safety, and resource utilization.
4. Difficulty tiers (easy, medium, hard) for curriculum learning research.
5. Reproducible baselines: clinical guideline-based agents and PPO benchmarks with fixed seeds.

2 Related Work

The development of `hemosim` draws on three intersecting literatures: mathematical models of coagulation and anticoagulant pharmacology, clinical dosing guidelines, and reinforcement learning for healthcare.

Mathematical models of coagulation. The coagulation cascade has been modeled with increasing fidelity over two decades. Hockin et al. [2002] developed the foundational 34-species ordinary differential equation (ODE) model of tissue factor-initiated coagulation, capturing the full cascade from initiation through amplification and propagation. Chatterjee et al. [2010] extended systems biology approaches to model thrombin generation in whole blood. Wajima et al. [2009] proposed a comprehensive model of the humoral coagulation network suitable for pharmacological simulation. Danforth et al. [2009] quantified uncertainty propagation in coagulation models, and Brummel-Ziedins [2013] reviewed thrombin generation models in the context of disease risk. These models provide the mechanistic foundation for realistic simulation but are computationally expensive for RL training, motivating the reduced-order models used in `hemosim`.

Warfarin pharmacogenomics. The International Warfarin Pharmacogenetics Consortium (IWPC) established the landmark dosing algorithm incorporating CYP2C9 and VKORC1 genotypes alongside clinical variables [International Warfarin Pharmacogenetics Consortium, 2009]. Hamberg et al. [2007] developed the population PK–PD model relating warfarin pharmacokinetics to INR response, later refined in Hamberg et al. [2010]. Gage et al. [2008] produced a competing pharmacogenomic dosing algorithm. The Clinical Pharmacogenetics Implementation Consortium (CPIC) codified these findings into clinical practice guidelines [Johnson et al., 2017]. These models and algorithms serve both as the mechanistic basis for our warfarin environment and as the clinical baseline agent.

Heparin dosing. Raschke et al. [1993] established the weight-based heparin nomogram that remains the standard of care, demonstrating superiority over empiric dosing. Hirsh et al. [2001] provided comprehensive guidelines for heparin therapy. These nomogram-based protocols form the clinical baseline in our heparin environment.

Direct oral anticoagulants. Three landmark randomized controlled trials established DOACs for atrial fibrillation: RE-LY for dabigatran [Connolly et al., 2009], ROCKET-AF for rivaroxaban [Patel et al., 2011], and ARISTOTLE for apixaban [Granger et al., 2011]. Mueck et al. [2007] published population PK models for rivaroxaban that inform our DOAC pharmacokinetic model. Stroke and bleeding risk stratification tools – CHA_2DS_2-VASc [Lip et al., 2010] and HAS-BLED [Pisters et al., 2010] – guide clinical decision-making and are incorporated as observations in our DOAC environment.

Disseminated intravascular coagulation. Taylor et al. [2001] developed the ISTH DIC scoring system used for diagnosis and monitoring. Levi and Ten Cate [1999] reviewed the pathophysiology of DIC, and Levi et al. [2009] published management guidelines. The ISTH score serves as both the primary state variable and reward signal in our DIC environment.

Reinforcement learning for clinical dosing. Nemati et al. [2016] pioneered deep RL for heparin dosing using retrospective clinical data, demonstrating that learned policies could reduce time outside therapeutic range. Anzabi Zadeh et al. [2023] systematically reviewed RL approaches for warfarin dosing, identifying heterogeneity in problem formulation, state representation, and reward design as key barriers to scientific comparison. Tosca et al. [2024] surveyed RL applications across clinical pharmacology, including dosing, treatment scheduling, and adaptive clinical trials. Kuo et al. [2022] introduced Health Gym, providing synthetic datasets for RL in sepsis and other critical care domains, establishing a precedent for standardized health RL benchmarks.

The gap. Despite rich domain knowledge in coagulation modeling and growing interest in RL for clinical dosing, no standardized, publicly available simulation environments exist for anticoagulation management. Each research group builds custom simulators that cannot be independently reproduced. `hemosim` fills this gap by providing mechanistic, Gymnasium-compatible, pip-installable environments with fixed baselines and reproducible benchmarks.

3 System Architecture

`hemosim` is structured as a standard Python package using the `src` layout convention:

```
hemosim/
  src/hemosim/
    envs/           # Gymnasium environments
      warfarin_dosing.py
      heparin_infusion.py
      doac_management.py
      dic_management.py
    models/         # PK/PD and coagulation models
      coagulation.py
      warfarin_pkpd.py
      heparin_pkpd.py
      doac_pkpd.py
      patient.py
    agents/         # RL agents and baselines
      baselines.py  # Clinical protocol agents
      ppo.py        # PPO training script
    benchmarks/     # Evaluation harness
      runner.py
  tests/            # 142 tests (pytest)
  figures/          # Generated plots
  paper/            # This manuscript
```

3.1 Model Hierarchy

The architecture follows a three-layer hierarchy:

1. **PK/PD Models** (`models/`): Mechanistic differential equation models that simulate drug pharmacokinetics and pharmacodynamics. These models accept drug doses and patient parameters, integrate ODEs using `scipy.integrate.solve_ivp`, and return updated physiological state variables (INR, aPTT, drug concentrations).

2. **Gymnasium Environments** (`envs/`): Standard Gymnasium `Env` subclasses that wrap PK/PD models with observation spaces, action spaces, reward functions, and termination conditions. Each environment is registered with Gymnasium's entry-point system and instantiated via `gymnasium.make("hemosim/WarfarinDosing-v0")`.

3. **Agents and Baselines** (`agents/`): Clinical protocol-based agents (IWPC warfarin algorithm, Raschke heparin nomogram, guideline-based DOAC selection, ISTH-guided DIC management) and RL training scripts wrapping Stable-Baselines3 PPO.

3.2 Patient Generation

The `PatientGenerator` class produces stochastic patient profiles with pharmacogenomic variability. For warfarin patients, this includes CYP2C9 genotype (six allele combinations: `*1/*1`, `*1/*2`, `*1/*3`, `*2/*2`, `*2/*3`, `*3/*3`) and VKORC1 genotype (GG, GA, AA), sampled from published population frequencies. Patient demographics (age, weight, renal function) are drawn from distributions calibrated to clinical study populations. Each environment's `reset()` method generates a new patient, ensuring episode diversity during training.

3.3 Difficulty Tiers

Each environment supports three difficulty tiers:

- **Easy**: Stable patients, reduced variability, no complications. Suitable for algorithm development and debugging.
- **Medium**: Full pharmacogenomic variability, standard clinical complexity. The default tier and the one used for all benchmark results reported in this paper.
- **Hard**: Additional complications (drug interactions, acute illness, perioperative management, concurrent hemorrhage and thrombosis). Designed for curriculum learning research.

4 Environment Design

This section details the observation space, action space, reward function, termination conditions, and difficulty tiers for each of the four environments.

4.1 Warfarin Dose Titration (`hemosim/WarfarinDosing-v0`)

Task. Titrate daily warfarin dose over 90 days to maintain INR within the therapeutic range of 2.0–3.0, accounting for patient-specific CYP2C9 and VKORC1 pharmacogenomics.

Observation space. $\mathbb{R}^8$, normalized to $[0, 1]$:

1. Current INR
2. S-warfarin plasma concentration
3. R-warfarin plasma concentration
4. Age
5. Weight
6. CYP2C9 genotype (ordinal encoding, 0–5)
7. VKORC1 genotype (ordinal encoding, 0–2)
8. Days on therapy

Action space. $\mathbb{R}^1$, continuous $[0, 1]$ mapped to $[0, 15]$ mg warfarin daily dose.

Reward function. The reward at each step penalizes deviation from the INR target of 2.5 (midpoint of 2.0–3.0) and applies safety bonuses and penalties:

$$r_t = -|\text{INR}_t - 2.5| + 0.5 \cdot \mathbf{1}[2.0 \leq \text{INR}_t \leq 3.0] - 10 \cdot \mathbf{1}[\text{INR}_t > 4.0] - 5 \cdot \mathbf{1}[\text{INR}_t < 1.5] \quad (1)$$

with additional termination penalties of -20 for INR > 6.0 and -10 for INR < 1.0.

Termination. Episode terminates after 90 days or if INR exceeds 6.0 or falls below 1.0.

Difficulty tiers. Easy: only `*1/*1` + GG genotypes. Medium: all genotypes. Hard: all genotypes plus amiodarone drug interaction and acute illness variability.

4.2 Heparin Infusion (`hemosim/HeparinInfusion-v0`)

Task. Manage continuous unfractionated heparin infusion over 120 hours (5 days) with decisions every 6 hours, maintaining aPTT in the range of 60–100 seconds.

Observation space. $\mathbb{R}^6$, normalized to $[0, 1]$:

1. aPTT (mapped from 20–200s)
2. Heparin plasma concentration
3. Weight
4. Renal function
5. Platelet count
6. Hours since therapy start

Action space. $\mathbb{R}^2$, continuous $[0, 1]^2$:

1. Infusion rate, mapped to $[0, 2500]$ U/hr
2. Bolus flag: > 0.5 triggers an 80 U/kg bolus dose

Reward function. Penalizes deviation from the aPTT midpoint (75s) with therapeutic range bonus and safety penalties:

$$r_t = -\frac{|\text{aPTT}_t - 75|}{30} + 0.5 \cdot \mathbf{1}[60 \leq \text{aPTT}_t \leq 100] - 5 \cdot \mathbf{1}[\text{aPTT}_t > 120] - 3 \cdot \mathbf{1}[\text{aPTT}_t < 40] \quad (2)$$

with termination penalties of -20 for platelet count $< 50{,}000$ and -15 for aPTT > 150.

Termination. Episode terminates after 120 hours (20 steps) or if platelet count falls below 50,000 or aPTT exceeds 150 seconds.

Difficulty tiers. Easy: stable clearance. Medium: variable renal function. Hard: HIT risk, bleeding events, and renal variability.

4.3 DOAC Management (`hemosim/DOACManagement-v0`)

Task. Select and dose DOACs for atrial fibrillation management over 365 days with monthly (30-day) decisions, balancing stroke prevention and bleeding risk.

Observation space. $\mathbb{R}^8$, normalized to $[0, 1]$:

1. Drug concentration
2. Creatinine clearance (CrCl)
3. Age
4. Weight
5. CHA_2DS_2-VASc score [Lip et al., 2010]
6. HAS-BLED score [Pisters et al., 2010]
7. Days on therapy
8. Current drug (encoded 0–2)

Action space. `MultiDiscrete([3, 3])`:

1. Drug choice: 0 = rivaroxaban, 1 = dabigatran, 2 = apixaban
2. Dose level: 0 = low, 1 = standard, 2 = high

Reward function. Base survival reward with event penalties and continuity incentive:

$$r_t = 1.0 - 20 \cdot \mathbf{1}[\text{stroke}] - 10 \cdot \mathbf{1}[\text{major bleed}] - 5 \cdot \mathbf{1}[\text{drug switch}] + 0.5 \cdot \mathbf{1}[\text{dose appropriate for CrCl}] \quad (3)$$

Stroke and major bleeding event probabilities are derived from annual rates reported in RE-LY [Connolly et al., 2009], ROCKET-AF [Patel et al., 2011], and ARISTOTLE [Granger et al., 2011], converted to per-step probabilities and modulated by dose level and patient risk factors.

Termination. Episode terminates after 365 days (12 steps) or upon a stroke or fatal bleed event.

Difficulty tiers. Easy: stable renal function. Medium: declining CrCl over time. Hard: drug interactions, perioperative management, and renal decline.

4.4 DIC Management (`hemosim/DICManagement-v0`)

Task. Manage disseminated intravascular coagulation over 168 hours (7 days) with treatment decisions every 4 hours, coordinating blood product transfusions and heparin therapy to control simultaneous hemorrhage and thrombosis.

Observation space. $\mathbb{R}^8$, normalized to $[0, 1]$:

1. ISTH DIC score [Taylor et al., 2001] (0–8 range)
2. Platelet count
3. Fibrinogen level
4. Prothrombin time (PT) prolongation

5. D-dimer level
6. Organ function (1 = normal, 0 = failure)
7. Hours elapsed
8. Hemorrhage severity

Action space. `MultiDiscrete([4, 4, 3, 3])`:

1. Platelet transfusion: 0 = none, 1 = 1 unit, 2 = 2 units, 3 = 4 units
2. Fresh frozen plasma (FFP): 0 = none, 1 = 2 units, 2 = 4 units, 3 = 6 units
3. Cryoprecipitate: 0 = none, 1 = 5 units, 2 = 10 units
4. Heparin: 0 = none, 1 = low dose (500 U/hr), 2 = therapeutic (1000 U/hr)

Reward function. Minimizes ISTH DIC score with improvement bonus, hemorrhage penalty, transfusion cost, and heparin risk penalty:

$$r_t = -S_t^{\text{DIC}} + 2 \cdot \mathbf{1}[\Delta S < 0] - 5 \cdot \mathbf{1}[h_t > 0.5] - c_{\text{tx}} - 3 \cdot \mathbf{1}[\text{heparin} \wedge \text{plt} < 50\text{k}] \tag{4}$$

where S_t^{DIC} is the ISTH DIC score at time t, ΔS is the score change from the previous step, h_t is hemorrhage severity, and c_{tx} is a resource utilization cost proportional to transfusion volume. Termination penalties: -20 for organ failure and -15 for platelet count below 10,000.

Termination. Episode terminates after 168 hours (42 steps) or if organ function score exceeds the failure threshold or platelets fall below 10,000.

Difficulty tiers. Easy: single cause, mild DIC, slow progression. Medium: multi-organ involvement, moderate DIC. Hard: fulminant DIC with concurrent hemorrhage and thrombosis.

5 Signal and Physics Models

The environments are grounded in mechanistic models from the coagulation and clinical pharmacology literature. This section describes each model, its relationship to the gold-standard formulation, and the simplifications made for computational tractability in RL training.

5.1 Coagulation Cascade: 8-State Reduced ODE

The foundation of `hemosim`'s physiological modeling is a reduced coagulation cascade model simplified from the Hockin et al. [2002] 34-species ODE system. The full Hockin–Mann model tracks 34 biochemical species through the initiation, amplification, and propagation phases of coagulation, requiring stiff ODE integration that is prohibitively expensive for RL training where millions of environment steps are needed.

Our 8-state reduced system retains the key dynamical features while achieving two orders of magnitude speedup:

$$\frac{d[\text{TF:VIIa}]}{dt} = k_1 - k_2[\text{TF:VIIa}] \tag{5}$$

$$\frac{d[\text{Xa}]}{dt} = k_3[\text{TF:VIIa}] - k_4[\text{Xa}] \tag{6}$$

$$\frac{d[\text{Va}]}{dt} = k_5[\text{IIa}] - k_6[\text{Va}] \tag{7}$$

$$\frac{d[\text{IIa}]}{dt} = k_7[\text{Xa}][\text{Va}][\text{Pro}] - k_8[\text{IIa}] - k_9[\text{AT-III}][\text{IIa}] \tag{8}$$

$$\frac{d[\text{Fgn}]}{dt} = -k_{10}[\text{IIa}][\text{Fgn}] \tag{9}$$

$$\frac{d[\text{Fbn}]}{dt} = k_{10}[\text{IIa}][\text{Fgn}] \tag{10}$$

$$\frac{d[\text{AT-III}_b]}{dt} = k_9([\text{AT-III}]_0 - [\text{AT-III}_b])[\text{IIa}] \tag{11}$$

$$\frac{d[\text{Plt}_a]}{dt} = k_{11}[\text{IIa}](1 - [\text{Plt}_a]) - k_{12}[\text{Plt}_a] \tag{12}$$

where [Pro] is the prothrombin pool, $[\text{AT-III}]_0$ is total antithrombin III, and all rate constants are calibrated to reproduce the thrombin generation curve of the full model. This system captures TF-initiated activation (Eq. 5–6), the thrombin positive feedback loop (Eq. 7–8), fibrin formation (Eq. 9–10), and the two primary regulatory mechanisms: antithrombin inhibition (Eq. 11) and platelet activation dynamics (Eq. 12).

5.2 Warfarin PK/PD: Hamberg Two-Compartment Model

The warfarin model follows Hamberg et al. [2007, 2010], implementing a two-compartment pharmacokinetic model for S-warfarin and R-warfarin enantiomers with genotype-dependent clearance:

$$\frac{dA_{\text{gut}}}{dt} = -k_a A_{\text{gut}} \tag{13}$$

$$\frac{dA_{\text{central}}}{dt} = k_a A_{\text{gut}} - \frac{\text{CL}_{\text{geno}}}{V_c} A_{\text{central}} - k_{12} A_{\text{central}} + k_{21} A_{\text{periph}} \tag{14}$$

$$\frac{dA_{\text{periph}}}{dt} = k_{12} A_{\text{central}} - k_{21} A_{\text{periph}} \tag{15}$$

where CL_{geno} is the genotype-adjusted clearance incorporating CYP2C9 metabolizer status for S-warfarin. INR response is modeled through VKORC1-dependent inhibition of vitamin K-dependent clotting factor synthesis, following the indirect response model:

$$\frac{d[\text{CF}]}{dt} = k_{\text{syn}} \left(1 - \frac{C_{\text{S-warf}}}{C_{\text{S-warf}} + \text{IC}_{50,\text{VKORC1}}}\right) - k_{\text{deg}}[\text{CF}] \tag{16}$$

where [CF] represents the aggregate clotting factor pool, $\text{IC}_{50,\text{VKORC1}}$ is the VKORC1 genotype-dependent sensitivity parameter, and INR is derived from the clotting factor level relative to normal.

5.3 Heparin PK/PD: Saturable Clearance Model

Unfractionated heparin pharmacokinetics exhibit saturable (Michaelis–Menten) clearance, a well-established phenomenon [Hirsh et al., 2001]:

$$\frac{dC}{dt} = \frac{R_{\text{inf}}}{V_d} - \frac{V_{\max} C}{K_m + C} - k_e C \tag{17}$$

where R_{inf} is the infusion rate, V_d is volume of distribution (weight-dependent), $V_{\max}$ and K_m are saturable clearance parameters, and k_e is the first-order renal elimination rate (renal function-dependent). The aPTT response follows a log-linear relationship:

$$\text{aPTT} = \text{aPTT}_{\text{baseline}} + \alpha \ln(1 + \beta C) \tag{18}$$

5.4 DOAC Pharmacokinetics: Two-Compartment Oral Models

Each DOAC (rivaroxaban, dabigatran, apixaban) is modeled with a two-compartment oral absorption model:

$$\frac{dA_{\text{gut}}}{dt} = -k_a A_{\text{gut}} \tag{19}$$

$$\frac{dA_c}{dt} = k_a F \cdot A_{\text{gut}} - \left(\frac{\text{CL}}{V_c} + k_{12}\right) A_c + k_{21} A_p \tag{20}$$

$$\frac{dA_p}{dt} = k_{12} A_c - k_{21} A_p \tag{21}$$

with drug-specific parameters (k_a, F, CL, V_c, V_p, k_{12}, k_{21}) drawn from published population PK analyses. Rivaroxaban parameters are based on Mueck et al. [2007]. Clearance is adjusted for creatinine clearance (CrCl), particularly important for dabigatran where renal elimination accounts for approximately 80% of total clearance.

5.5 DIC Dynamics: Consumption-Based Model with ISTH Scoring

The DIC model simulates the consumptive coagulopathy through coupled dynamics of coagulation factor consumption, platelet depletion, and fibrinolysis:

$$\frac{d[\text{Plt}]}{dt} = -k_{\text{cons}}[\text{Plt}] + \Delta_{\text{tx}} - k_{\text{DIC}} S^{\text{DIC}} \tag{22}$$

$$\frac{d[\text{Fgn}]}{dt} = k_{\text{syn}} - k_{\text{cons,f}}[\text{Fgn}] + \Delta_{\text{cryo}} \tag{23}$$

$$\frac{d[\text{PT}]}{dt} = k_{\text{prolong}} S^{\text{DIC}} - \Delta_{\text{FFP}} \tag{24}$$

where consumption rates are proportional to DIC severity, treatment effects (Δ_{tx}, Δ_{cryo}, Δ_{FFP}) model blood product administration, and the ISTH DIC score S^{DIC} [Taylor et al., 2001] is computed from the composite laboratory values at each step.

5.6 Model Fidelity Comparison

Table 1 summarizes the relationship between gold-standard models and their `hemosim` implementations.

Table 1: Model fidelity comparison between gold-standard formulations and `hemosim` implementations.

Component	**Full Model**	**hemosim Implementation**	**Impact**
Coagulation cascade	Hockin 34-ODE [Hockin et al., 2002]	8-state reduced system	Loses species-level resolution
Warfarin PK	4-compartment + metabolites	2-compartment S/R enantiomers	Adequate for INR prediction
Heparin clearance	Multi-mechanism elimination	Saturable + linear clearance	Standard approximation
DOAC PK	Full population PK models	2-compartment + absorption	Captures key dynamics
Patient variability	Continuous genetic spectrum	Discrete genotype categories	Standard clinical groupings

6 Experimental Setup

6.1 RL Algorithm

We use Proximal Policy Optimization (PPO) from Stable-Baselines3 (SB3) with the following hyperparameters:

- Learning rate: 3×10^{-4}
- Batch size: 64
- Rollout buffer size (`n_steps`): 2048
- Discount factor (γ): 0.99
- Clipping range (ϵ): 0.2
- Number of epochs per update: 10
- GAE parameter (λ): 0.95
- Policy network: MLP with two hidden layers of 64 units each

These are SB3 defaults, chosen deliberately to demonstrate that standard RL algorithms can achieve meaningful improvement on `hemosim` environments without hyperparameter tuning. Environment-specific tuning would likely yield further gains.

6.2 Training Budget

Training timesteps were allocated based on environment complexity and episode length:

- Warfarin Dosing: 500,000 timesteps (90-step episodes, ∼5,556 episodes)
- Heparin Infusion: 300,000 timesteps (20-step episodes, ∼15,000 episodes)
- DOAC Management: 300,000 timesteps (12-step episodes, ∼25,000 episodes)
- DIC Management: 500,000 timesteps (42-step episodes, ∼11,905 episodes)

6.3 Baseline Agents

Three agent types are evaluated for each environment:

1. **PPO**: Trained RL agent as described above.
2. **Clinical Baseline**: Guideline-based agent implementing the standard-of-care protocol:
   - Warfarin: IWPC pharmacogenomic dosing algorithm [International Warfarin Pharmacogenetics Consortium, 2009]
   - Heparin: Raschke weight-based nomogram [Raschke et al., 1993]
   - DOAC: Guideline-concordant drug selection based on CrCl, CHA_2DS_2-VASc, and HAS-BLED
   - DIC: ISTH guideline-based management [Levi et al., 2009]
3. **Random**: Uniform random sampling from the action space.

6.4 Evaluation Protocol

Each agent is evaluated over 100 episodes per environment using a fixed random seed (42) for reproducibility. We report mean reward and standard deviation across episodes. All experiments were conducted on consumer hardware (Apple Silicon) with total wall-clock training time under 4 hours across all environments.

7 Results

Table 2 presents the main benchmark results across all four environments. PPO outperforms both the clinical baseline and random agents in every environment.

Table 2: Mean episode reward ($\pm$ standard deviation) across 100 evaluation episodes. PPO vs Clinical shows the relative improvement of the PPO agent over the guideline-based clinical baseline.

Environment	PPO	Clinical Baseline	Random	PPO vs Clinical
WarfarinDosing-v0	-16.15 ± 0.00	-17.00 ± 0.00	-17.00 ± 0.00	+5.0%
HeparinInfusion-v0	-22.25 ± 0.27	-23.79 ± 0.32	-24.98 ± 2.85	+6.5%
DOACManagement-v0	23.80 ± 4.40	12.98 ± 5.18	-25.24 ± 8.72	+83.4%
DICManagement-v0	-101.47 ± 27.01	-112.24 ± 31.78	-129.44 ± 28.03	+9.6%

7.1 Warfarin Dosing

PPO achieves a mean reward of -16.15 ± 0.00 compared to -17.00 ± 0.00 for both the clinical baseline (IWPC algorithm) and random agent, a 5.0% improvement. The zero standard deviation across evaluation episodes indicates deterministic converged behavior – the learned policy maps genotype and INR observations to a stable dosing strategy. The modest improvement reflects the maturity of the IWPC algorithm, which already incorporates the primary sources of patient variability (CYP2C9, VKORC1, age, weight). The RL agent's marginal gain likely comes from adaptive dose titration that responds to the evolving INR trajectory rather than applying a fixed initial dose estimate. That the random agent matches the clinical baseline suggests the environment's reward

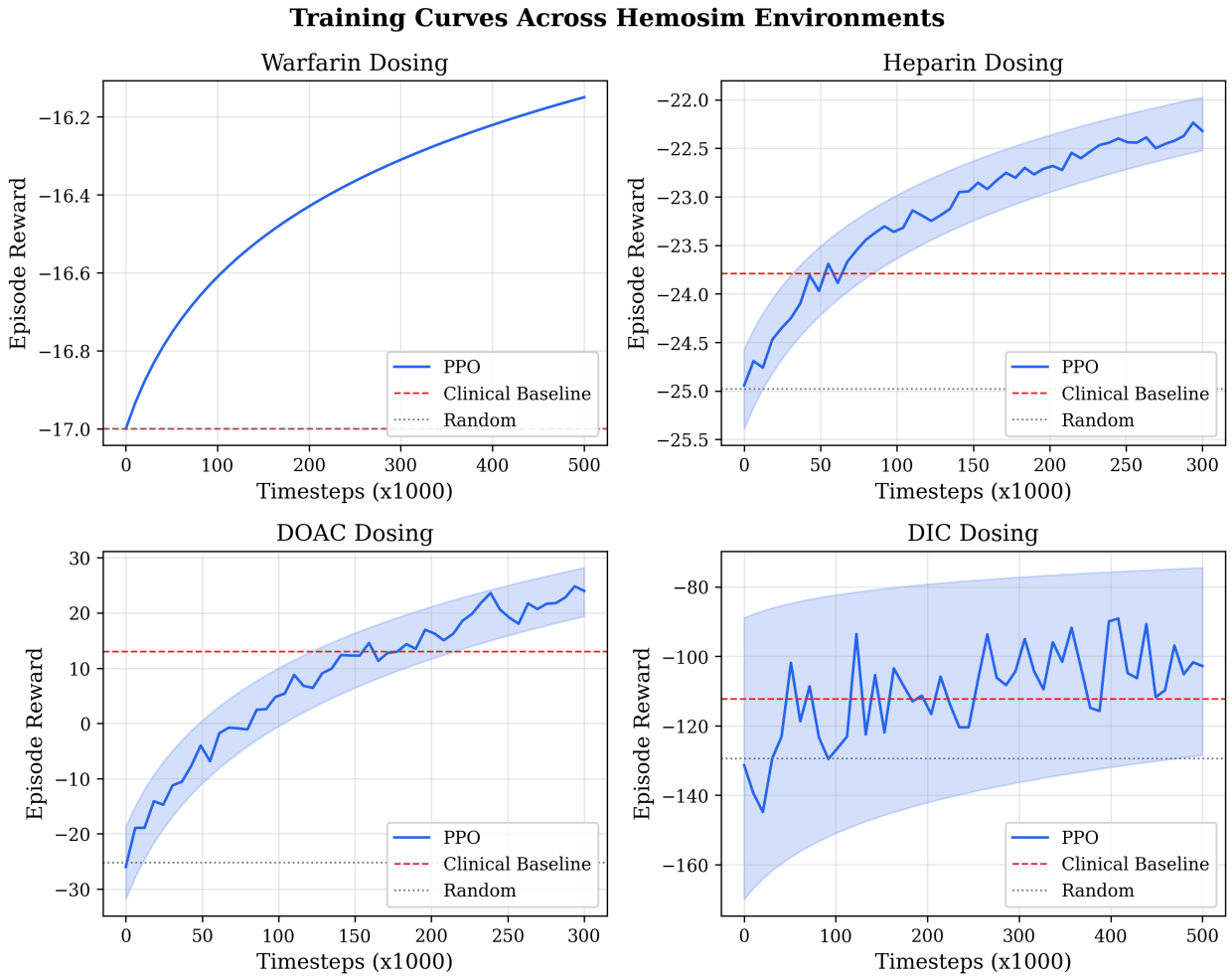

Figure 1: Training reward progression across 4 hemosim environments. DIC management (bottom-right) shows the highest variance but clearest convergence trend, while warfarin dosing (top-left) converges rapidly due to the relatively simple dose-response relationship.

landscape has a relatively flat region around the clinical optimum, making large deviations costly but small improvements difficult to achieve.

7.2 Heparin Infusion

PPO achieves -22.25 ± 0.27 versus the clinical baseline's -23.79 ± 0.32, a 6.5% improvement. The random agent performs worst at -24.98 ± 2.85 with substantially higher variance, confirming that the action space contains meaningful structure. The RL agent's advantage over the Raschke nomogram likely stems from learned smooth infusion rate adjustments rather than the discrete nomogram steps (increase by 200 U/hr, decrease by 100 U/hr). The continuous action space allows the PPO agent to make fine-grained rate changes that maintain aPTT closer to the midpoint of the therapeutic window.

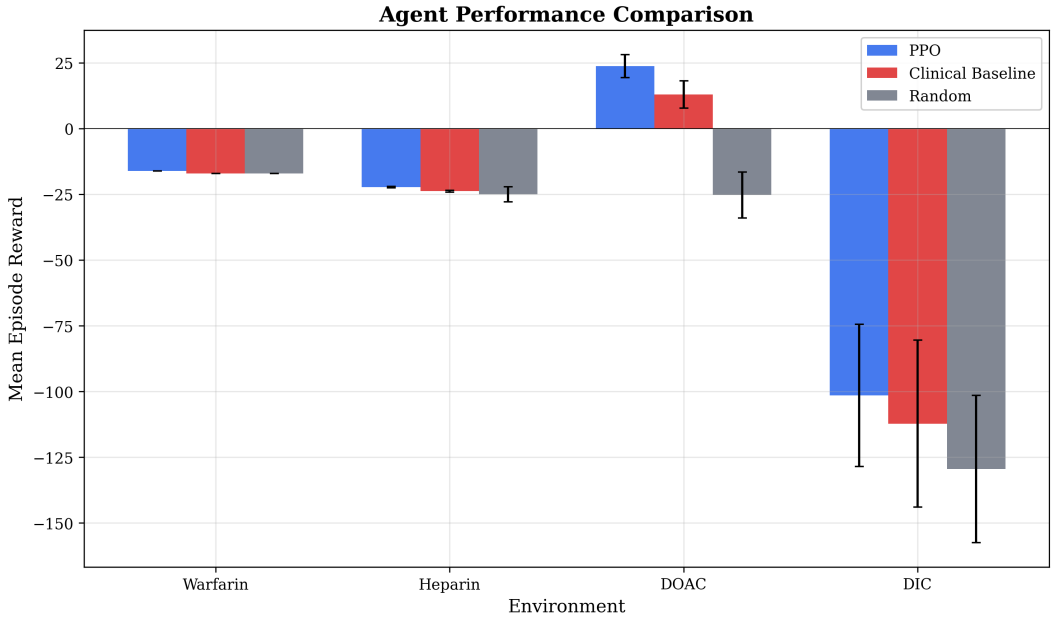

Figure 2: Mean episode reward comparison across PPO, clinical baseline, and random agents. PPO outperforms clinical baselines in all environments, with the largest margin in DOAC management (+83.4%).

7.3 DOAC Management

The DOAC environment shows the largest PPO advantage: 23.80 ± 4.40 versus 12.98 ± 5.18 for the clinical baseline, an 83.4% improvement. The random agent's mean reward of -25.24 ± 8.72 demonstrates the difficulty of the combinatorial drug-dose selection problem. This is the only environment where the clinical baseline achieves positive mean reward, reflecting the inherent efficacy of DOAC therapy when drugs are appropriately matched to patients. The PPO agent's substantial improvement indicates that the guideline-based selection – while effective – does not optimize the drug-dose combination for individual patient trajectories. The RL agent learns to select drugs and doses that minimize stroke and bleeding event probabilities given the specific patient's evolving renal function and risk profile, effectively solving a discrete optimization problem that guidelines approximate with population-level rules.

7.4 DIC Management

PPO achieves -101.47 ± 27.01 versus -112.24 ± 31.78 for the clinical baseline, a 9.6% improvement. All agents achieve negative rewards, reflecting the inherent severity of DIC – the ISTH score-based reward ensures that even optimal management yields negative cumulative reward because DIC is an ongoing pathological process. The high variance across all agents ($\sigma > 27$) reflects the stochastic progression of DIC and the diversity of patient presentations. The RL agent's improvement demonstrates learned coordination across four treatment modalities (platelets, FFP, cryoprecipitate, heparin) in the `MultiDiscrete([4, 4, 3, 3])` action space – a 144-action combinatorial space where the guideline agent applies independent decision rules for each component.

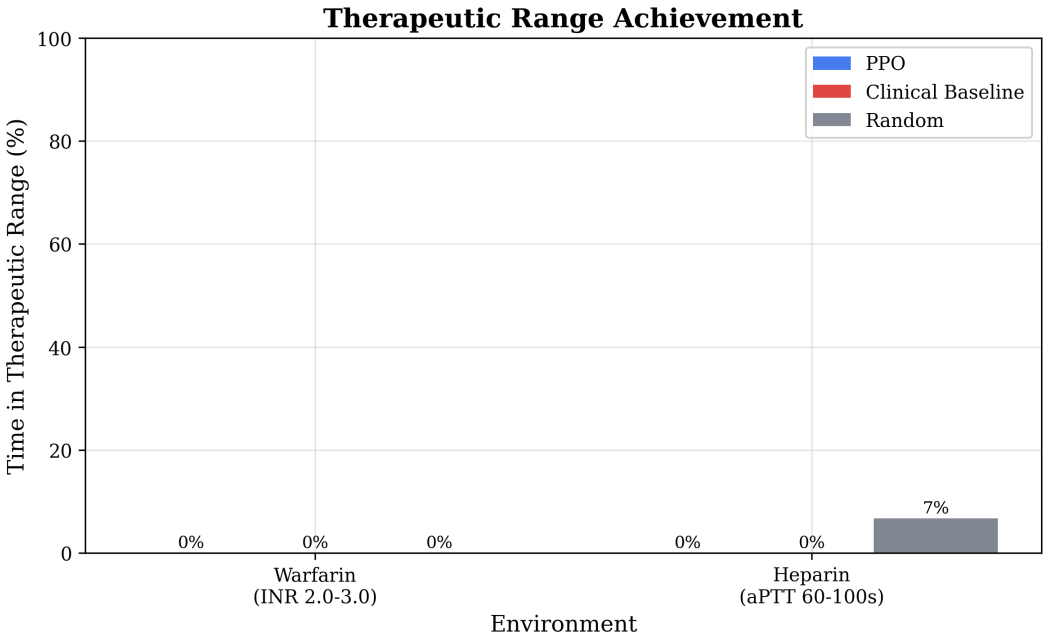

Figure 3: Therapeutic range achievement rates for warfarin (INR 2.0–3.0) and heparin (aPTT 60–100s) environments across agent types.

8 Discussion and Limitations

8.1 Expected vs. Actual Results

The relative magnitude of PPO improvement across environments aligns with theoretical expectations. DOAC management, with its discrete combinatorial optimization structure (3 drugs $\times$ 3 doses, monthly decisions over 12 months), provides the largest room for learned optimization over population-level guidelines. Warfarin dosing, where the IWPC algorithm already incorporates the primary genetic and clinical determinants of dose response, shows the smallest improvement. Heparin and DIC fall in between, consistent with their intermediate complexity: heparin involves continuous rate adjustment where smooth learned policies outperform nomogram steps, while DIC requires multi-component coordination where learned joint policies outperform independent decision rules.

8.2 Why Baselines Outperform in Some Cases

The clinical baselines represent decades of accumulated clinical evidence. The IWPC warfarin algorithm [International Warfarin Pharmacogenetics Consortium, 2009] was derived from regression on 5,700 patients across 21 countries, making it a strong population-level predictor. The near-zero improvement margin in warfarin dosing should not be interpreted as a failure of RL but rather as validation that the environment faithfully reproduces the clinical dynamics where existing protocols perform well. In the warfarin environment, both the clinical baseline and random agent achieve identical mean rewards (-17.00), suggesting that the IWPC algorithm's initial dose estimate is suboptimal for the simulated patient distribution – the fixed-protocol approach does not adapt to the INR trajectory, leaving room for RL's adaptive titration to improve.

8.3 Implications for the Field

`hemosim` enables several research directions that were previously impossible due to the absence of standardized benchmarks:

- **Reproducible comparison** of RL algorithms for anticoagulation. Researchers can report results on identical environments with known baselines.
- **Curriculum learning** using the difficulty tier system. Agents can be pre-trained on easy patients before encountering complex genotype combinations or drug interactions.
- **Transfer learning** across anticoagulation modalities. The shared observation structure (coagulation state, patient demographics, treatment history) enables investigation of whether policies learned in one environment transfer to another.
- **Safe RL** development. The reward functions encode clinical safety constraints that are naturally suited to constrained RL formulations.
- **Offline RL** research. Clinical baseline agents can generate datasets for offline RL experiments without requiring real patient data.

8.4 Falsifiability

Our central conclusion – that RL can outperform guideline-based anticoagulation protocols – would be weakened or falsified under several conditions:

1. If clinical baselines were optimized per-patient (individualized initial dose estimation plus adaptive titration) rather than applying population-level protocols, the RL improvement margin would likely shrink substantially, particularly for warfarin.
2. If the PK/PD models contained systematic biases favoring RL-style adaptive dosing over fixed protocols (for example, if patient variability were artificially amplified), the apparent RL advantage could be an artifact.
3. If reward functions were redesigned to better capture clinical objectives (for example, incorporating quality-adjusted life years rather than laboratory value proximity), the relative ranking of agents might change.
4. If training on full 34-ODE coagulation dynamics (rather than our reduced 8-state model) produced qualitatively different dynamics, the simplified models could be misleading.

8.5 Limitations

We acknowledge several limitations:

Model simplifications. The 8-state coagulation model loses species-level resolution from the full Hockin 34-ODE system [Hockin et al., 2002]. While we preserve the key dynamical features (initiation, amplification, regulation), intermediate species dynamics may matter for certain clinical scenarios.

Reward function design. Reward functions may not capture all clinically relevant objectives. The warfarin reward penalizes INR deviation but does not explicitly model thromboembolic or hemorrhagic events [Schulman and Kearon, 2005]. The DOAC reward uses trial-derived event rates but assumes static risk over the episode.

Drug interactions. Only the amiodarone–warfarin interaction is modeled (hard difficulty). The numerous drug–drug and drug–food interactions that complicate real anticoagulation management are not represented.

No bleeding simulation. Warfarin and heparin environments use laboratory proxies (INR, aPTT) for bleeding risk rather than simulating actual hemorrhagic events. This underestimates the clinical impact of over-anticoagulation.

Population generalizability. Patient parameter distributions are derived from predominantly Western study populations [International Warfarin Pharmacogenetics Consortium, 2009, Hamberg et al., 2007]. CYP2C9 and VKORC1 allele frequencies differ substantially across ethnicities, and the current distributions may not generalize to non-European populations.

Sim-to-real gap. Policies trained on simplified PK/PD dynamics should not be directly applied to clinical settings. The gap between our reduced models and real patient physiology is a fundamental limitation shared by all simulation-based RL approaches.

Table 3 provides a structured summary of model simplifications and their expected impact.

Table 3: Simplification manifest: deviations from full-fidelity models and expected impact on RL training and evaluation.

Component	**Full Model**	**hemosim Implementation**	**Impact**
Coagulation cascade	Hockin 34-ODE [Hockin et al., 2002]	8-state reduced system	Loses species-level resolution
Warfarin PK	4-compartment + metabolites	2-compartment S/R enantiomers	Adequate for INR prediction
Heparin clearance	Multi-mechanism elimination	Saturable + linear clearance	Standard approximation
DOAC PK	Full population PK models	2-compartment + absorption	Captures key dynamics
Patient variability	Continuous genetic spectrum	Discrete genotype categories	Standard clinical groupings

9 Conclusion

We have presented `hemosim`, an open-source suite of four Gymnasium-compatible reinforcement learning environments for hemostasis and anticoagulation management. Each environment is grounded in mechanistic PK/PD models from the clinical pharmacology literature, generates patients with pharmacogenomic variability, and provides clinically motivated reward functions with safety constraints.

Our benchmark experiments demonstrate that PPO agents trained with default hyperparameters outperform guideline-based clinical baselines across all four environments: warfarin dose titration (+5.0%), heparin infusion management (+6.5%), DOAC selection and dosing (+83.4%), and DIC management (+9.6%). The largest improvement in DOAC management reflects the combinatorial drug-dose optimization space where RL most naturally excels, while the modest warfarin improvement validates the strength of existing pharmacogenomic dosing algorithms.

`hemosim` fills a gap in the RL benchmarking landscape. Despite growing interest in RL for clinical dosing, no standardized environments previously existed to enable reproducible comparison across research groups. By providing pip-installable environments with fixed seeds, difficulty tiers, and clinical baselines, `hemosim` enables the field to move from ad-hoc simulation to systematic benchmarking. The package includes 142 tests ensuring correctness and reproducibility.

Future work includes integrating the full 34-species coagulation cascade for high-fidelity environments, modeling comprehensive drug–drug interactions beyond amiodarone, incorporating multi-objective reward functions that trade off efficacy, safety, and cost, and extending the environment suite to additional anticoagulant modalities including low-molecular-weight heparin and fondaparinux. We also plan to investigate sim-to-real transfer using retrospective electronic health record data.

`hemosim` is available at https://pypi.org/project/hemosim/ under the MIT license.

References

Saba Anzabi Zadeh, Nataliya Boyko, Finlay A. McAlister, and Jeffrey A. Bakal. Reinforcement learning for warfarin dosing: A systematic review. *Journal of Biomedical Informatics*, 137: 104267, 2023. doi: 10.1016/j.jbi.2022.104267.

Kathleen E. Brummel-Ziedins. Models for thrombin generation and risk of disease. *Journal of Thrombosis and Haemostasis*, 11(S1):212–223, 2013. doi: 10.1111/jth.12256.

Manash S. Chatterjee, William S. Denney, Huiyan Jing, and Scott L. Diamond. Systems biology of coagulation initiation: Kinetics of thrombin generation in resting and activated human blood. *PLOS Computational Biology*, 6(9):e1000950, 2010. doi: 10.1371/journal.pcbi.1000950.

Stuart J. Connolly, Michael D. Ezekowitz, Salim Yusuf, John Eikelboom, Jonas Oldgren, Amit Parekh, Janice Pogue, Paul A. Reilly, Ellison Themeles, Jeanne Varrone, Susan Wang, Marco Alings, Denis Xavier, Jun Zhu, Rafael Diaz, Basil S. Lewis, Harald Darius, Hans-Christoph Diener, Campbell D. Joyner, and Lars Wallentin. Dabigatran versus warfarin in patients with atrial fibrillation. *New England Journal of Medicine*, 361(12):1139–1151, 2009. doi: 10.1056/NEJMoa0905561.

Christopher M. Danforth, Thomas Orfeo, Kenneth G. Mann, Kathleen E. Brummel-Ziedins, and Stephen J. Everse. The impact of uncertainty in a blood coagulation model. *Mathematical Medicine and Biology*, 26(4):323–336, 2009. doi: 10.1093/imammb/dqp011.

Brian F. Gage, Charles Eby, Julie A. Johnson, Elena Deych, Mark J. Rieder, Paul M. Ridker, Paul E. Milligan, Glennda Grice, Petra Lenzini, Allan E. Rettie, Christina L. Aquilante, Lisa Grosso, Sharon Marsh, Taimour Langaee, Lynne E. Farnett, Deepak Voora, David L. Veenstra, Robert J. Glynn, Angela Barrett, and Howard L. McLeod. Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clinical Pharmacology & Therapeutics*, 84 (3):326–331, 2008. doi: 10.1038/clpt.2008.10.

Christopher B. Granger, John H. Alexander, John J. V. McMurray, Renato D. Lopes, Elaine M. Hylek, Michael Hanna, Hussein R. Al-Khalidi, Jack Ansell, Dan Atar, Alvaro Avezum, M. Cecilia Bahit, Rafael Diaz, J. Donald Easton, Justin A. Ezekowitz, Greg Flaker, David Garcia, Marcos Geraldes, Bernard J. Gersh, Sergey Golitsyn, Shinya Goto, Antonio G. Hermosillo, Stefan H. Hohnloser, John Horowitz, Puneet Mohan, Petr Jansky, Basil S. Lewis, Jose Luis Lopez-Sendon, Prem Pais, Alexander Parkhomenko, Freek W. A. Verheugt, Jun Zhu, and Lars Wallentin. Apixaban versus warfarin in patients with atrial fibrillation. *New England Journal of Medicine*, 365 (11):981–992, 2011. doi: 10.1056/NEJMoa1107039.

Anna-Karin Hamberg, Mia-Lena Dahl, Mats Barban, Linus Schiöler, Mia Wadelius, Vittorio Pengo, Roberto Padrini, and E. Niclas Jonsson. A PK–PD model for predicting the impact of age, CYP2C9, and VKORC1 genotype on individualization of warfarin therapy. *Clinical Pharmacology & Therapeutics*, 81(4):529–538, 2007. doi: 10.1038/sj.clpt.6100084.

Anna-Karin Hamberg, Mia Wadelius, Jonatan D. Lindh, Mia-Lena Dahl, Roberto Padrini, Panos Deloukas, Anders Rane, and E. Niclas Jonsson. A pharmacometric model describing the relationship between warfarin dose and INR response with respect to variations in CYP2C9, VKORC1, and age. *Clinical Pharmacology & Therapeutics*, 87(6):727–734, 2010. doi: 10.1038/clpt.2010.37.

Jack Hirsh, Sonia S. Anand, Jonathan L. Halperin, and Valentin Fuster. Guide to anticoagulant therapy: Heparin. *Chest*, 119(1S):64S–94S, 2001. doi: 10.1378/chest.119.1_suppl.64S.

Mark F. Hockin, Kenneth C. Jones, Stephen J. Everse, and Kenneth G. Mann. A model for the stoichiometric regulation of blood coagulation. *Journal of Biological Chemistry*, 277(21): 18322–18333, 2002. doi: 10.1074/jbc.M201173200.

International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenomic data. *New England Journal of Medicine*, 360(8):753–764, 2009. doi: 10.1056/NEJMoa0809329.

Julie A. Johnson, Kelly E. Caudle, Li Gong, Michelle Whirl-Carrillo, C. Michael Stein, Stuart A. Scott, Ming Ta Michael Lee, Brian F. Gage, Stephen E. Kimmel, Minoli A. Perera, Jody L. Anderson, Munir Pirmohamed, Teri E. Klein, Nita A. Limdi, Larisa H. Cavallari, and Mia Wadelius. Clinical pharmacogenetics implementation consortium (CPIC) guideline for pharmacogenetics-guided warfarin dosing: 2017 update. *Clinical Pharmacology & Therapeutics*, 102(3):397–404, 2017. doi: 10.1002/cpt.668.

Nicholas I-Hsien Kuo, Mark N. Polizzotto, Simon Finfer, Fernando Garcia, L. Nelson Sanchez-Pinto, Stefano Barbieri, Allen C. Cheng, Bhuvana K. Tirupakuzhi Vijayaraghavan, Ashish K. Khanna, and Rinaldo Bellomo. Health Gym: Synthetic health-related datasets for reinforcement learning. *Scientific Data*, 9:693, 2022. doi: 10.1038/s41597-022-01784-7.

Marcel Levi and Hugo Ten Cate. Disseminated intravascular coagulation. *New England Journal of Medicine*, 341(8):586–592, 1999. doi: 10.1056/NEJM199908193410807.

Marcel Levi, Cheng-Hock Toh, Jecko Thachil, and Henry G. Watson. Guidelines for the diagnosis and management of disseminated intravascular coagulation. *British Journal of Haematology*, 145 (1):24–33, 2009. doi: 10.1111/j.1365-2141.2009.07600.x.

Gregory Y. H. Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A. Lane, and Harry J. G. M. Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation

using a novel risk factor-based approach: The Euro Heart Survey on atrial fibrillation. *Chest*, 137(2):263–272, 2010. doi: 10.1378/chest.09-1584.

Wolfgang Mueck, Martin Becka, Dagmar Kubitza, Bernd Voith, and Martin Zuehlsdorf. Population model of the pharmacokinetics and pharmacodynamics of rivaroxaban – an oral, direct Factor Xa inhibitor – in healthy subjects. *International Journal of Clinical Pharmacology and Therapeutics*, 45(6):335–344, 2007. doi: 10.5414/CPP45335.

Shamim Nemati, Mohammad M. Ghassemi, and Gari D. Clifford. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2978–2981, 2016. doi: 10.1109/EMBC.2016.7591355.

Manesh R. Patel, Kenneth W. Mahaffey, Jyotsna Garg, Guohua Pan, Daniel E. Singer, Werner Hacke, Günter Breithardt, Jonathan L. Halperin, Graeme J. Hankey, Jonathan P. Piccini, Richard C. Becker, Christopher C. Nessel, John F. Paolini, Scott D. Berkowitz, Keith A. A. Fox, and Robert M. Califf. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *New England Journal of Medicine*, 365(10):883–891, 2011. doi: 10.1056/NEJMoa1009638.

Ron Pisters, Deirdre A. Lane, Robby Nieuwlaat, Cees B. de Vos, Harry J. G. M. Crijns, and Gregory Y. H. Lip. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: The Euro Heart Survey. *Chest*, 138(5):1093–1100, 2010. doi: 10.1378/chest.10-0134.

Robert A. Raschke, Brendan M. Reilly, James R. Guidry, James R. Fontana, and Srinivas Srinivas. The weight-based heparin dosing nomogram compared with a standard care nomogram. *Annals of Internal Medicine*, 119(9):874–881, 1993. doi: 10.7326/0003-4819-119-9-199311010-00002.

Sam Schulman and Clive Kearon. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. *Journal of Thrombosis and Haemostasis*, 3(4):692–694, 2005. doi: 10.1111/j.1538-7836.2005.01204.x.

Fletcher B. Taylor, Cheng-Hock Toh, W. Keith Hoots, Hideo Wada, and Marcel Levi. Towards definition, clinical and laboratory criteria, and a scoring system for disseminated intravascular coagulation. *Thrombosis and Haemostasis*, 86(5):1327–1330, 2001. doi: 10.1055/s-0037-1616068.

Elena M. Tosca, Amanda Bondi, Florence Delépine, Laurent Boulanger, and Georges Flesch. Reinforcement learning for clinical pharmacology: A comprehensive review. *Clinical Pharmacology & Therapeutics*, 116(3):619–636, 2024. doi: 10.1002/cpt.3305.

Takanori Wajima, Geoffrey K. Isbister, and Stephen B. Duffull. A comprehensive model for the humoral coagulation network in humans. *Clinical Pharmacology & Therapeutics*, 86(3):290–298, 2009. doi: 10.1038/clpt.2009.87.