POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Exploring the Potential of Reinforcement Learning for Automated Anesthesia Control

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Manuela Merlo, 222181

Advisor:
Prof. Francesco Trovò

Co-advisors:
Prof. Alberto Maria Metelli
Dr. Emiliano Tognoli

Academic year:
2023-2024

Abstract:
Each year, more than 300 million surgical procedures worldwide require general anesthesia. However, despite its complexity, this process is still managed manually. Automated closed-loop systems have been investigated to enhance patient well-being and have demonstrated improvements in stability and patient safety. This study presents three Reinforcement Learning (RL) agents for anesthesia control during the maintenance phase, trained and evaluated using a patient simulator we designed to provide an in-simulation environment. The first agent mimics the policy of a Proportional-Integral-Derivative (PID) controller, adjusting the hypnotic agent to regulate the Depth of Hypnosis (DoH) under a fixed hypnotic-to-opioid ratio. The second agent incorporates patient demographic data and removes the fixed ratio assumption. The third agent extends the second agent's objective by monitoring hemodynamic stability. Using standard control metrics, the RL agents are compared to the PID controller. The first agent demonstrates superior performance in controlling the sedative and the DoH, while the second and third agents perform comparably to the PID controller while managing more complex tasks. Although not explicitly trained to handle disturbances, all agents can compensate for DoH perturbations but struggle with Mean Arterial Pressure deviations. The explainability analysis shows that the policies of the second and third agents align with the Pharmacokinetic and Pharmacodynamic models on which they were trained. These findings suggest that RL offers a promising approach to personalized anesthesia control.

Key-words: automatic anesthesia control, reinforcement learning, pharmacokinetic-pharmacodynamic modeling, patient simulator software, patient-specific treatment

1. Introduction

Each year, more than 300 million surgical procedures worldwide require general anesthesia [30]. This medically induced, reversible state enables patients to undergo major surgical procedures safely [74]. It involves the administration of anesthetic drugs to induce and maintain unconsciousness, suppress pain, and relax muscles. In the literature, some of the most frequently mentioned drugs for closed-loop control of anesthesia include propofol as a hypnotic agent, remifentanil as an analgesic, and rocuronium as a neuromuscular blocking agent [39, 66].

Currently, anesthesiologists rely on monitoring systems to assess patients' states and adjust drug administration manually or with the help of Target-Controlled Infusion (TCI) systems. Therefore, in recent years, different technologies have been introduced in operating rooms to monitor and optimize anesthesia delivery [34]. However, despite significant advances, general anesthesia remains a challenging task.

One major factor is the considerable inter- and intra-patient variability [15, 33, 41, 92]. Individual differences in Pharmacokinetics (PK) and Pharmacodynamics (PD) cause different patient reactions to anesthetic agents, making it difficult to determine the right dosage of drugs to achieve adequate anesthesia. Additionally, the complexity is worsened by human factors that increase the likelihood of making mistakes [20].

Because errors in the administration of anesthetic drugs can lead to intraoperative awareness [59], impact postoperative outcomes, and even cause mortality [55], there is a growing need to develop closed-loop controllers that reduce anesthesiologists' cognitive burden and help them administer anesthesia in a patient-specific manner. Automated systems, such as Proportional-Integral-Derivative (PID) [7] control, and Model Predictive Control (MPC) [48], have been shown to improve stability and patient safety in anesthesia management [16]. PID controllers, widely used for their simplicity and robustness, have been successfully applied in systems like the one developed by Heusden et al. [35] for propofol-remifentanil co-administration. The results demonstrated robust stability and improved disturbance rejection. However, PID's limitations include fixed parameters that prevent adaptability to patient variability and limited control actions. In contrast, systems developed by Pawłowski et al. [63] and Sandre-Hernandez et al. [67] leverage the MPC approach to predict patient responses and optimize control strategies accordingly. MPC has the advantage of handling complex and non-linear systems and improving stability. However, challenges remain, such as computational complexity and dependence on model accuracy.

Reinforcement Learning (RL) [79] offers a model-free and adaptive approach for patient-specific treatments. Studies such as those by Padmanabhan et al. [60] and Schamberg et al. [68] have demonstrated RL's ability to control anesthesia dosing, effectively handling system uncertainties while aligning with current best practices. However, current systems focus on single-drug control regulating one physiological state, limiting their potential to exploit drug interactions and manage multiple parameters simultaneously.

This work presents three RL-based controllers, all aimed at managing the maintenance phase, which is the longest and accounts for the highest percentage of human errors [20] while omitting the effects of external factors or complications. The first agent shares the same objective and enhances the performance of a PID controller, which serves as a reference. The PID controller, inspired by the study in [52], is designed to optimize propofol infusion rates regulating the NeuroWave Index (WAV_{cns}), which is used to assess Depth of Hypnosis (DoH), using a fixed propofol-to-remifentanil ratio. The state representation of the first agent includes proportional, integral, and derivative errors relative to the target DoH value. The second agent incorporates patient demographics to co-administer propofol and remifentanil, enabling personalized anesthesia delivery. Finally, the third agent builds on this approach by also regulating hemodynamic stability. The proposed approaches utilize continuous state and action spaces, allowing for finer control over drug administration.

Yet, despite its potential, applying RL to anesthesia control presents significant challenges, especially concerning the need for large volumes of high-quality training data. Heusden et al. [35] evaluated the reliability of the PID controller through clinical experiments. However, RL agents learn by exploring different scenarios and observing the consequences of various actions. In anesthesia, replicating such scenarios on real patients is unsafe and ethically prohibited. Other studies [60, 68, 93, 94], to overcome the limitations of clinical trials, employ simple population-based Pharmacokinetic (PK)-Pharmacodynamic (PD) models or patient simulators such as the one developed by Ionescu et al. [39].

To address this challenge, this work introduces the Anesthesia Response Simulator (AReS), a novel comprehensive tool conceived to mimic the response of patients undergoing anesthesia. The simulator is designed based on clinical practice, literature models, and medical experience. AReS incorporates some of the most common drugs and measured outputs used during surgical operations. To validate it as a tool for designing decision-making systems, we empirically demonstrate that the multi-variable simulation on a set of patients is consistent with the corresponding signals measured during anesthesia. In addition, we analyze the sensitivity of this simulator to different settings. Thanks to AReS, we train our agents in a simulated environment, allowing them to explore a wide range of randomized demographic data and action combinations. Additionally, we use it to optimize the parameters of the PID controller with demographic data from 44 real individuals to regulate the induction and maintenance phases.

When using a simulator, an RL agent learns by interacting with PK-PD models, similar to those used in MPC. Consequently, in both cases, the effectiveness of the controller in real-world situations is limited by the accuracy of these representations. However, RL has the advantage of being able to generalize beyond the fixed model. Furthermore, we use explainable techniques, namely SHapley Additive exPlanations (SHAP) [47] values, to show that RL agents can learn to exploit demographic data to tailor actions for each specific patient in a manner consistent with the PK-PD models employed. As mentioned earlier, we compare the performance of our agents with that of the PID controller, using the same 44 patients for which the PID parameters were optimized, employing standard metrics for closed-loop anesthesia control.

This work serves as a **proof of concept**, demonstrating the feasibility of using RL with a complex model in simulation. The next step would be to collect real-world data to evaluate whether these techniques can be effectively applied. With this data, we could assess existing strategies and leverage RL to uncover complex patterns, enabling personalized treatment strategies that exceed the capabilities of traditional models.
In summary, the main contributions of this thesis are as follows:

- Development of a novel patient anesthesia simulator that provides an in-simulation environment for training and testing closed-loop anesthesia controllers.
- Implementation of three RL anesthesia controllers with progressively more complex objective. The first controls the DoH and propofol, the second adds remifentanil regulation influenced by demographic data, and the third incorporates hemodynamic regulation.
- Comparison of simulator results with clinical data and analysis of its sensitivity to different combinations of model configurations.
- Performance evaluation of RL against a PID controller using standard performance metrics.
- Analysis of RL-based decision-making using explainable techniques.

The remainder of this thesis is structured as follows: Section 2 provides an overview of general anesthesia, current administration techniques, and associated challenges. Section 3 reviews existing closed-loop control strategies and simulators and their limitations. Section 4 defines the control objectives and formulates them as Markov Decision Processes (MDPs). Section 5 details the design of the Anesthesia Response Simulator. Section 6 presents the reinforcement learning framework, including the description of the environment, the training algorithm, and the hyperparameter configuration. Section 7 describes the implementation of the PID controller used as a reference. Section 8 describes the experimental assessment of the simulator and RL agents. Finally, Section 9 summarizes the key findings and discusses limitations and potential future research directions.

2. Medical Background

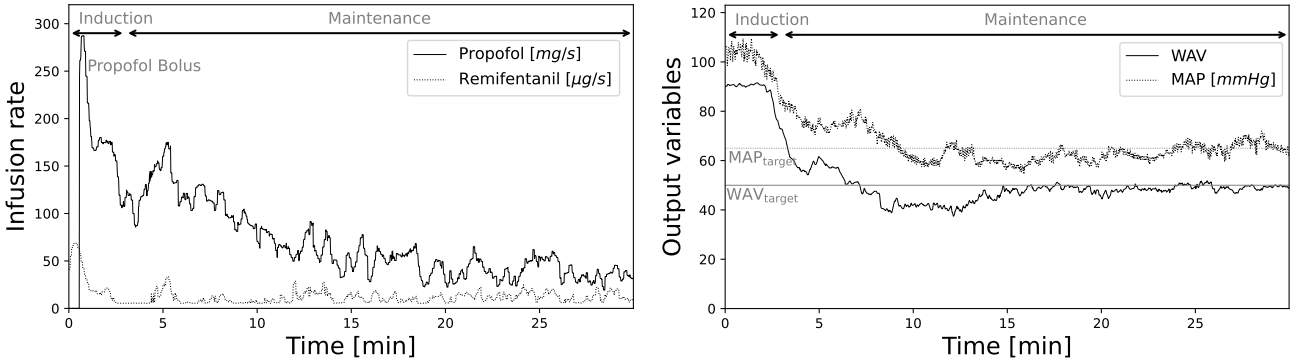
General Anesthesia (GA) can be defined as a medically induced reversible state that allows patients to undergo surgery [74]. This procedure involves administering drugs to induce amnesia, analgesia, hypnosis, and akinesia. Specifically, we focus on Total Intravenous Anesthesia (TIVA), where anesthesia is maintained exclusively with intravenous drugs [2]. The main category of anesthetic agents used in TIVA includes hypnotics for amnesia and opioids for analgesia. Neuromuscular blocking agents are used when muscle relaxation is needed. Propofol, remifentanil, and rocuronium, a hypnotic, opioid, and neuromuscular blocking agent, respectively, have been cited in the clinical and biomedical engineering literature as the ones that are more suitable for continuous infusion used in closed-loop control frameworks due to their fast onset and recovery times [39, 66].

2.1. Phases of General Anesthesia

General anesthesia comprises three distinct phases: induction, maintenance, and emergence [17]. Figure 1 illustrates the median infusion rates of propofol and remifentanil (Figure 1a) and the corresponding WAV_{cns} and Mean Arterial Pressure (MAP) values (Figure 1b) for 10 patients, based on data from [89]. In both subfigures, the transition from induction to maintenance is visible. *Induction* is the initial phase in which the patient is transitioned smoothly from consciousness to unconsciousness while maintaining hemodynamic stability. As shown in Figure 1a, this interval is characterized by an intravenous bolus of an anesthetic agent. During these measurements, opioids were co-administered to reduce pain perception. Neuromuscular blocking agents may also be used, particularly for procedures requiring intubation [17]. *Maintenance* begins after induction and continues until the end of surgery. In this case, the objective is to sustain the desired level of anesthesia while compensating for external disturbances, such as surgical stimuli or physiological changes like hypovolemia. In the case of TIVA, anesthesia is maintained through a continuous infusion of hypnotics and opioids. *Emergence* marks the end of anesthesia. The goal of this final phase is to restore consciousness by suspending drug administration, allowing plasma drug concentrations to decrease while continuously monitoring the patient's physiological status. The emergence time can vary depending on the patient's metabolism and the total amount of drugs administered.

2.2. Monitoring the Physiological State of the Patient

Anesthesiologists rely on various physiological indices to assess the DoH, hemodynamic stability, and level of akinesia in a patient. Based on these assessments, they can adjust drug administrations during the surgical procedure accordingly.
The Bispectral Index (BIS) index [43], an indicator derived from electroencephalogram (EEG) signals, is one



(a) (b)

Figure 1: General Anesthesia Phases.

of the most commonly used to assess the DoH. This parameter provides a statistically derived value that ranges from 0 (complete suppression of brain activity) to 100 (full wakefulness) [32]. Despite widespread use, Bibian et al. [10] demonstrated that the BIS monitor exhibits an average delay of approximately 5 seconds in displaying the BIS signal. To address this issue, a new index called WAV_{cns} [12] has been developed. Unlike BIS, the monitor that tracks WAV_{cns} [11] provides real-time monitoring without delay, offering a more responsive assessment. During induction and maintenance, a value of BIS or WAV_{cns} in the $[40, 60]$ range is generally considered safe to prevent awareness and the risk of delayed recovery and cognitive dysfunction [32].
Considering hemodynamic stability, MAP is a key parameter, as it is the main determinant of organ perfusion. Sustained MAP below $60-70$ mmHg has been associated with increased risks of acute kidney injury, myocardial injury, and eventually death. Currently, there is no universal definition of intraoperative hypotension, making it difficult to define a reproducible decision-making process [26]. However, in the event of hypotension, it is important to assess the patient's Cardiac Output (CO), Stroke Volume (SV), Heart Rate (HR), and systemic vascular resistance (SVR) to identify the underlying cause [51]. In addition, evaluating CO and SV helps guide fluid administration to restore adequate tissue perfusion. Fellahi et al. [26] recommend a target MAP of at least 65 mmHg or within $10-20\%$ of preoperative baseline values. These findings are supported by Meng et al. [51], who explains that for patients with normal baseline blood pressure, the objective is to maintain a MAP between approximately $[65, 95]$ mmHg to ensure adequate organ perfusion.
Lastly, Neuromuscular Blockade (NMB) is usually monitored using two techniques: Train-Of-Four (TOF) [3] stimulation and Post-Tetanic Count (PTC) [85]. TOF stimulation is generally used when moderate paralysis is required. The latter consists of four electrical stimuli at 2 Hz. Train-of-Four Ratio (TOFR) is defined as the ratio of the amplitudes of the fourth contraction (T4) and the first contraction (T1). When neuromuscular blocking agents take effect, the first response that decreases is the fourth T4 stimulus, followed by T3, T2, and T1 if complete paralysis is achieved. This decrement in twitch height is known as *fade*. During recovery, contractions reappear in reverse order, with T1 returning first and T4 last [50]. When $\text{TOF} = 0$, the neuromuscular block is too deep to be evaluated with TOF, so PTC stimulation is used [85]. PTC is defined as the number of responses to a sustained tetanic stimulation at 50 Hz applied for 5 seconds, followed by a pause of 3 seconds, and then single-twitch stimuli delivered every second for 10 up to 20 seconds. The target PTC or TOF responses depend on the level of muscle relaxation required for the procedure. Tsai and Mangar [80] categorized neuromuscular blockade into different levels: complete ($\text{PTC} = 0$), profound ($\text{PTC} = 1-2$), deep-moderate ($\text{TOF} = 0, \text{PTC} \geq 3$), moderate ($\text{TOF} = 1-3$), shallow ($\text{TOFR} \leq 0.9$), acceptable recovery ($\text{TOFR} \geq 0.9$), and full recovery ($\text{TOFR} = 1$).

2.3. Pharmacokinetics and Pharmacodynamics Models

Anesthesiologists and the systems they use to administer anesthesia rely on Pharmacokinetic and Pharmacodynamic models to guide drug delivery. Pharmacokinetics describes how a drug moves through the body, including absorption, distribution, metabolism, and elimination. PK models predict drug concentrations in different tissues over time based on mathematical models. A common approach is to represent the body using a mammillary compartment model, where the number of compartments can vary depending on the drug [76]. Specifically, three-compartment models (Figure 2) are commonly used to represent the pharmacokinetics of propofol and remifentanil. In this characterization, the body is divided into a central compartment, which represents the site where the drug is infused (the bloodstream or plasma), and two peripheral compartments that correspond to moderately perfused tissues (fast peripheral compartment), such as muscle, and poorly perfused tissues (slow peripheral compartment), such as fat. The transfer of the drug between these compartments is

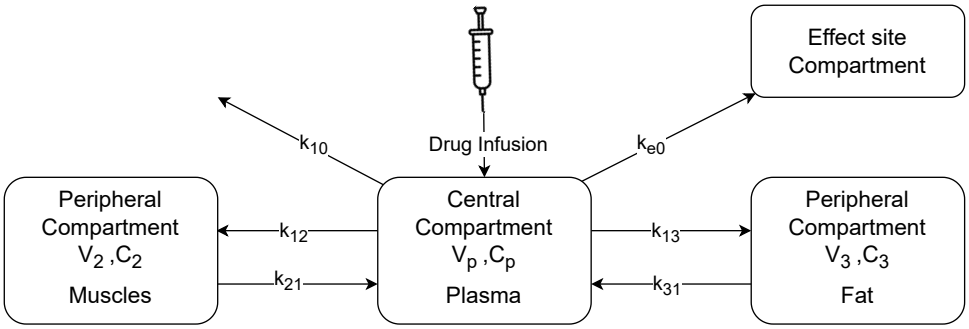


Figure 2: Schematic View of a three-compartment PK-PD model.

governed by rate constants k_{ij} that are influenced by patients' demographic parameters, including age, weight, height, Body Mass Index (BMI) [96], Lean Body Mass (LBM), and gender (e.g., Eleveld model [25]).
The temporal evolution of drug concentration in each compartment in a three-compartment model is described by the following system of differential equations [76]:

$$\frac{dC_p}{dt} = -(k_{10} + k_{12} + k_{13})C_p + k_{21}C_2 + k_{31}C_3 + \frac{I(t)}{V_1}, \tag{1}$$

$$\frac{dC_2}{dt} = k_{12}C_p - k_{21}C_2, \tag{2}$$

$$\frac{dC_3}{dt} = k_{13}C_p - k_{31}C_3, \tag{3}$$

where C_p, C_2, and C_3 denote the drug concentrations in the central, fast, and slow peripheral compartments, respectively. $I(t)$ represents the input infusion rate. k_{e0} describes the rate at which a drug moves from the central compartment to the effect site (the organ where the drug exercises therapeutic action). k_{10} represents the elimination rate constant. The exchange of the drug between the central and fast peripheral compartments is described by the rate constants k_{12} and k_{21}. In contrast, the exchange between the central and slow peripheral compartments is defined by k_{13} and k_{31}.
While Pharmacokinetics studies how drug concentrations evolve in the body, Pharmacodynamic models define the relationship between plasma concentrations and the drug effect. PD models usually consist of two main components: a linear part describing how the drug reaches equilibrium between the plasma and the effect site and a non-linear component quantifying the drug's therapeutic effect based on its concentration at the effect site. The relationship between the effect-site concentration C_e (or, in some cases, plasma concentration C_p) and the observed drug effect E is typically described by a variation of the Hill equation [76]:

$$E(t) = E_0 + (E_{\max} - E_0)\frac{C(t)^\gamma}{EC_{50}^\gamma + C(t)^\gamma}. \tag{4}$$

This equation determines the drug effect based on the plasma or effect-site concentration $C(t)$, accounting for E_0 and $E_{\max}$ that represents the baseline effect and the maximum effect, respectively. EC_{50} is the concentration required to produce 50% of the maximum effect, and $\gamma \geq 1$ is the steepness of the dose-response curve.
We note that since the parameters of population-based PK and PD models are derived from population-level data, they may not always provide an accurate prediction of an individual patient's response.

2.4. Target Controlled Infusion Systems

TCI systems are computer-controlled pumps widely used in clinical practice to achieve a user-defined target drug concentration in a specific body compartment [77]. TCI systems work as open-loop control systems, meaning they do not receive real-time feedback from the patient's actual plasma or effect-site drug concentrations. As illustrated in Figure 3, a TCI system consists of three components: an interface used by a decision-maker, a software, and an infusion pump [1]. Through the interface, an anesthesiologist or a decision-making system sets the target concentration $T_{\text{target}}(t)$ for the desired compartment, which can be either the plasma or the effect site. The target concentration is adjusted based on the prediction of the drug concentration in all compartments, $C_{\text{predicted}}(t)$, and monitoring signals such as WAV_{cns}, hemodynamic variables, or NMB. The TCI software consists of an algorithm and an internal PK-PD model. The internal PK-PD model simulates drug concentration in all compartments, $C_{\text{predicted}}(t)$, and at each time step, the algorithm computes the next drug infusion rate, $u(t+1)$, such that the prediction of the drug concentration in the desired compartment reaches $T_{\text{target}}(t)$ as

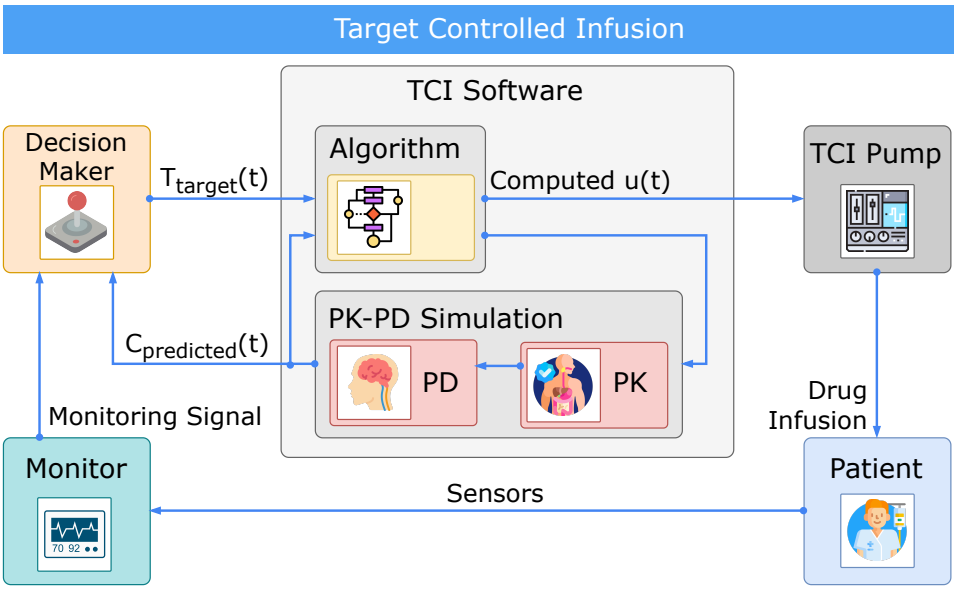


Figure 3: Schematic View of a TCI system.

quickly as possible. Note that since the TCI is using approximate PK-PD models, the predicted concentrations, $C_{\text{predicted}}(t)$, may be different from those of the patient. Therefore, the response might differ from that desired by the user, requiring repeated adjustment of the TCI target $T_{\text{target}}(t)$ throughout the process.

2.5. Challenges

General anesthesia remains a complex task despite significant advances in administration techniques and monitoring technologies. Challenges include administering an optimal dose that ensures adequate anesthesia while avoiding both underdosing and overdosing. Insufficient anesthesia can lead to intraoperative awareness, which has been associated with an increased risk of post-traumatic stress disorder (PTSD) [59]. Conversely, excessive anesthesia has been associated with poor postoperative outcomes, including postoperative delirium and increased mortality [27, 55].
Significant inter- and intra-patient variability in drug metabolism and response represents another remarkable concern. Indeed, the effects of anesthetic agents can vary considerably among patients due to factors such as age, weight, and gender [15, 33, 92]. Elderly patients, for example, usually require lower doses of anesthetics due to greater drug sensitivity [92]. Intra-patient variability is equally problematic, as changes in hemodynamics and tissue perfusion during surgery, such as those caused by blood loss, can alter the pharmacokinetics of drugs [41].
Finally, human factors and potential errors contribute to risks to patient safety. Cooper et al. [20] applied a modified version of critical incident analysis to examine preventable anesthetic incidents. Among the 359 descriptions of preventable incidents collected during the study, 82% involved human errors, with frequent issues including respiratory circuit disconnections and drug syringe exchange errors. Results indicate that human errors often result from insufficient experience, unfamiliarity with equipment, and poor communication. In addition, fatigue and distractions, whether due to the environment or the multitasking nature of the work, increase the likelihood of errors. Haste has been identified as the fourth most frequent contributing factor.
These challenges underscore the need for decision support systems that assist anesthesiologists by automating repetitive tasks and tailoring treatment to the individual patient.

3. Related Works

Automated systems have been shown to enhance stability and patient safety compared to manual administration [16]. Various control strategies have been investigated, with PID [7] control and MPC [48] being among the most widely used approaches.
PID controllers have been studied and implemented in the anesthesia automation literature because of their simplicity, robustness, and low computational requirements. Heusden et al. [35] developed a Multiple-Input, Single-Output (MISO) control system for propofol-remifentanil co-administration to regulate the DoH. Evaluation of the system in 80 clinical cases demonstrated that the controller provided robust stability and improved disturbance rejection. However, the system did not handle hemodynamic stability, which is essential for the

comprehensive management of general anesthesia. Similarly, Merigo et al. [52] presented a control scheme for BIS regulation using propofol-remifentanil co-administration. Their findings showed that the proposed solution met clinical specifications during the induction and maintenance phases. Furthermore, Monte Carlo simulations demonstrated that the system was robust to inter- and intra-patient variability. However, in addition to not considering hemodynamic variables, the approach relied on a fixed propofol-to-remifentanil ratio. This constraint may prevent the system from finding optimal dosing strategies, limiting the ability to tailor anesthesia to individual patient responses and potentially using higher drug dosages than needed.

Although PID control is advantageous and works well in many scenarios, it has some limitations. Fixed control parameters limit adaptability to patient variability, making generalization to different patient populations difficult. Additionally, designing PID-based systems that effectively control highly non-linear physiological systems is challenging. The complexity of the control task increases further when considering multiple drugs.

MPC represents a more advanced option that exploits models to predict patient responses over a defined time horizon and optimizes the control strategy accordingly. Pawłowski et al. [63] designed an MPC system to control both propofol and remifentanil, employing a non-linear MISO model to predict their combined effect on BIS through a generalized MPC algorithm. However, like [52], this system relied on a fixed drug ratio, limiting patient-specific adaptability. Later, Pawlowski et al. [62] developed an MPC strategy using a Generalized Predictive Control (GPC) algorithm to regulate BIS, achieving the target in both the induction and maintenance phases while controlling propofol. However, this system did not include remifentanil administration. Further improvements were made by Sandre-Hernandez et al. [67], who removed the fixed drug ratio constraint and achieved stable BIS control in the induction and maintenance phases. However, hemodynamic regulation is still not addressed. Despite the advantages of MPC, several challenges remain, including computational complexity and model dependency. Indeed, mismatches between predicted and actual patient responses can degrade control performance.

Due to these challenges, research has explored RL as a more adaptive, model-free solution that can tailor patient-specific treatments. Padmanabhan et al. [60] proposed a deep RL-based agent that learned to control both the depth of hypnosis and MAP using propofol. Simulation results showed that the proposed RL control strategy was promising for designing closed-loop controllers and was robust to system uncertainties. Later, Schamberg et al. [68] proposed the first fully continuous deep RL algorithm for automating anesthetic drug dosing. Different agents were trained to control loss of unconsciousness using propofol. These models significantly outperformed a PID controller while providing recommendations consistent with best practices in anesthesia care. Calvi et al. [18] employed Q-learning to control propofol infusion rates and regulate the depth of hypnosis. The model learned a policy aligned with literature findings during the induction and maintenance phases. It could handle non-linearities arising from pharmacodynamics, noise effects on the BIS signal, intra- and inter-patient parameter variations, unknown surgical disturbances, and the anesthesiologist's presence in the loop. Later, Yun et al. [93] investigated a Hierarchical Deep Reinforcement Learning assistant system for propofol and remifentanil co-administration to optimize BIS control. The quantitative evaluation suggested that the infusion system was a potentially effective strategy for anesthesia.

Despite the promising results obtained using RL, existing implementations have mostly focused on regulating a single anesthetic agent and/or controlling only one physiological parameter rather than jointly managing multiple drugs and patient responses. Yet, single-drug control systems overlook the synergy between anesthetic drugs like propofol and remifentanil [13], reducing the effectiveness of automated control. In addition, most implemented controllers define a state that depends only on the error relative to the output variables considered and/or values derived from them.

Heusden et al. [35] evaluated closed-loop controllers through clinical experiments. However, RL agents learn by exploring different scenarios and observing the results. In anesthesia, replicating extreme conditions such as hypotension or deep hypnosis on real patients is unsafe and ethically prohibited.

To overcome the limitations of clinical trials, other studies have employed simulation experiments to design and assess decision-making systems for anesthesia. The robustness of various control strategies, including RL [18], has been validated using the patient simulator developed by Ionescu et al. [39].

To the best of our knowledge, the scientific literature provides two multivariable simulators for the anesthesia field: one by Ionescu et al. [39] developed in Matlab, and one by Aubouin-Pairault et al. [8] developed in Python. They consider different control variables and output measurement signals, specified in Table 1. However, the PK-PD models used therein do not simulate hemodynamic variables, such as HR and SV, since they do not use physiological models. Moreover, the two studies considered a linear interaction between drug responses to predict hemodynamic variables, i.e., between CO and MAP. At the same time, it has been shown that such an interaction is more complex in real-world patients. Regarding the physiological response to surgical stimuli and blood loss, they only allow the user to simulate disturbances with a fixed dynamic and intensity without considering subject-specific characteristics or current drug concentrations. Consequently, this approach restricts the user's ability to investigate the effects of different surgical treatments that may also depend on the patient's physiology.

Table 1: Comparison of Input and Output Variables across Ionescu et al. [39], Aubouin-Pairault et al. [8], and AReS simulators.

Input	Ionescu	Aubouin-Pairault	AReS (ours)
Propofol	✓	✓	✓
Remifentanil	✓	✓	✓
Norepinephrine	✗	✓	✓
Atracurium	✓	✗	✗
Rocuronium	✗	✓	✓
Sodium Nitroprusside	✓	✗	✗
Dopamine	✓	✗	✗
Output	Ionescu	Aubouin-Pairault	AReS (ours)
BIS	✓	✓	✓
WAV_{cns}	✗	✗	✓
HR	✗	✗	✓
SV	✗	✗	✓
CO	✓	✓	✓
MAP	✓	✓	✓
RASS	✓	✗	✗
Tolerance Of Laryngoscopy (TOL)	✗	✓	✗
NMB	✓	✗	✓

4. Problem Formulation

This work focuses on the maintenance phase of anesthesia, the longest and most error-prone phase in clinical practice [20], while excluding the impact of external factors or complications on the patient's condition. To investigate the potential of RL in anesthesia control, we formulate our problem as a progressive learning framework, introducing increasing levels of complexity at each stage. The problems are defined as Markov Decision Processes [65]. An MDP is a mathematical framework for sequential decision-making problems in which an agent interacts with an environment. Sutton and Barto [79] describe Reinforcement Learning as process where, at each step, the agent receives a state $s \in \mathcal{S}$, selects an action $a \in \mathcal{A}$ following its policy $\pi : \mathcal{S} \times \mathcal{A}$, and receives an *immediate reward* r and the new state of the environment $s' \in \mathcal{S}$ (see Figure 4).
An MDP can be formally defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the set of states, and $\mathcal{A}$ is the set of available actions to the agent. $P(s'|s, a)$ is the transition probability function, which defines the probability of reaching state $s' \in \mathcal{S}$ when action taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$:

$$P(s'|s, a) = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a).$$

The reward function $R(s, a)$ represents the expected reward obtained when taking action $a \in \mathcal{A}$ from state $s \in \mathcal{S}$. Finally, $\gamma \in [0, 1]$ is the discount factor, determining future rewards' importance. A value of $\gamma = 0$ makes the agent short-sighted, only caring about immediate rewards, while a value close to 1 makes it far-sighted. Consequently, the agent may sacrifice the immediate reward to gain long-term rewards. The agent behavior is fully defined by a policy $\pi(a|s)$, which is a probability distribution over actions $a \in \mathcal{A}$ given a state $s \in \mathcal{S}$:

$$\pi(a|s) = \mathbb{P}[a_t = a|s_t = s].$$

Given a policy π, we define the value function $V^\pi(s)$ of a state s, as the expected return when starting in s and following π thereafter:

$$V^\pi(s) \doteq \mathbb{E}_\pi[G_t|s_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|s_t = s\right], \quad \forall s \in \mathcal{S}.$$

Here, R_{t+k+1} represent the reward received at time step $t + k + 1$ and γ is the discount factor. However, for control purposes, it is often easier to consider the action-value function $Q^\pi(s, a)$ which represents the expected

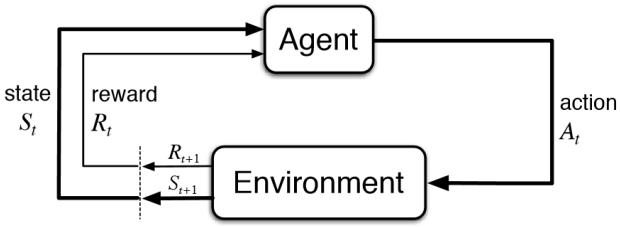


Figure 4: The Agent-Environment interaction in Reinforcement Learning (from [79]).

return starting from state s, taking action a and then following policy π. It can be written as:

$$Q^{\pi}(s,a) \doteq \mathbb{E}_{\pi}\left[G_t | s_t = s, a_t = a\right] = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s, a_t = a\right], \quad \forall s \in \mathcal{S}.$$

The main goal of an RL agent is to learn the *optimal policy* π^* that maximizes the expected return starting from any state $s \in \mathcal{S}$. The optimal state-value function $V^*(s)$ represents the maximum achievable value for each state:

$$V^*(s) \doteq \max_{\pi} V^{\pi}(s), \quad \forall s \in \mathcal{S}.$$

This is obtained when following the optimal policy π^*. Similarly, the optimal action-value function $Q^*(s,a)$ is the maximum achievable expected return starting from s, taking action a and then following the optimal policy π^*:

$$Q^*(s,a) \doteq \max_{\pi} Q^{\pi}(s,a), \quad \forall s \in \mathcal{S} \text{ and } \forall a \in \mathcal{A}.$$

In particular, the optimal *deterministic* policy μ^* is obtained from the optimal action-value function $Q^*(s,a)$ by selecting the action that maximizes $Q^*(s,a)$ for each state:

$$\mu^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a) \quad \forall s \in \mathcal{S}.$$

In the following, we detail the MDP formulation for each of the RL agents developed in this work.

4.1. RL Agent 1: Enhancing PID with Propofol Control

The first RL agent is designed to enhance the performance of the PID controller explained in Section 7 by learning an improved policy for propofol administration while keeping remifentanil at a fixed ratio ($u_{remi} = u_{prop}/2$). The problem is formulated as an MDP with a continuous state space $\mathcal{S}$ defined as:

$$s = (e_P, e_I, e_D)^T \in \mathbb{R}^3,$$

where e_P, e_I, e_D represent the proportional, integral, and derivative errors, respectively, at time step t. These errors for state $s_t = (e_P(t), e_I(t), e_D(t))$ are defined as:

$$e_P(t) = \frac{\text{WAV}(t) - 50}{50}, \quad e_I(t) = \sum_{i=0}^{t} e_P(i), \quad e_D(t) = \frac{e_P(t) - e_P(t-1)}{2}.$$

The action space consists of a continuous signal representing the infusion rate of propofol. The action range spans from -1 to 1, where a value of -1 indicates no infusion, and a value of 1 corresponds to the maximum infusion rate of 2 mg/s. The reward function R uses the absolute error to speed up learning and avoid excessively small gradients near the target, penalizing deviations from the ideal WAV_{cns}:

$$R(s,a) = -|\text{WAV}_{error}| = -\left|\frac{\text{WAV} - 50}{50}\right|.$$

We set the discount factor to $\gamma = 0.99$. Although a larger value might be preferable for long-term optimization, the higher complexity of the task requires a balance between long-term planning and efficient learning.

4.2. RL Agent 2: Removing the Fixed Ratio Constraint on Remifentanil

In the second approach, we remove the fixed drug ratio constraint and allow the agent to co-administer propofol and remifentanil freely. Additionally, we introduce patient demographics into the state representation to make the agent tailor drug administration to individual patient characteristics. A 1-unit error in an elderly or obese patient may require different management compared to a younger or anorexic individual. We expand the state space to include both demographic data and the physiological state:

$$s = (s_1, s_2, s_3, s_4, \mathrm{WAV}_{\mathrm{error}})^T \in \mathbb{R}^5,$$

where the first four components represent age s_1, height s_2, weight s_3, and gender s_4. The gender value s_4 is binary, where $s_4 = 1$ for female and $s_4 = 0$ for male. The action space consists of two continuous signals $(a_{\mathrm{prop}}, a_{\mathrm{remi}})$ representing propofol and remifentanil infusion rates. As with the previous MDP, the action range spans from -1 to 1, with the maximum infusion rate for remifentanil being 0.5 μg/s. The reward function is designed to encourage precise hypnosis control while trying to minimize drug administration:

$$R(s, a) = -|\mathrm{WAV}_{\mathrm{error}}| - 0.5\ (a_{\mathrm{prop}}^2 + a_{\mathrm{remi}}^2).$$

We penalize the drug volume to ensure safety and improve convergence speed. We maintain the discount factor $\gamma = 0.99$.

4.3. RL Agent 3: Integrating Hemodynamic Stability Control

Finally, we extend the control objective beyond hypnosis regulation to include hemodynamic stability. Since MAP and HR are interrelated (see Equation 7), we choose to control only MAP while monitoring its indirect effects on HR. We expand the state space as follows:

$$s = (s_1, s_2, s_3, s_4, \mathrm{WAV}_{\mathrm{error}}, \mathrm{MAP}_{\mathrm{error}})^T \in \mathbb{R}^6.$$

The action space remains the same, with continuous control over propofol and remifentanil infusion rates. The literature defines a safe range for MAP as 65 to 95 mmHg (see Section 2). However, Wanner et al. [86] suggests that strictly maintaining a MAP ≥ 75 mmHg does not provide significant benefits over the standard approach (targeting MAP≥ 65 mmHg). To account for this and reduce the complexity of the task, we introduce a dead zone in the range $[65, 75]$ mmHg, where the penalty for MAP-related errors is set to zero in the reward function. Outside this range, the MAP error is computed as follows:

$$\mathrm{MAP}_{\mathrm{error}} = \begin{cases} \frac{65-\mathrm{MAP}}{65}, & \text{if MAP} < 65 \\ \frac{\mathrm{MAP}-75}{75}, & \text{if MAP} > 75 \end{cases}.$$

The final reward function accounts for both hypnosis and hemodynamic stability while minimizing drug usage:

$$R(s, a) = -|\mathrm{WAV}_{\mathrm{error}}| - \mathrm{MAP}_{\mathrm{error}} - 0.5\ (a_{\mathrm{prop}}^2 + a_{\mathrm{remi}}^2).$$

The discount factor is still kept as $\gamma = 0.99$.

5. Simulator Development

Training RL agents for anesthesia automation requires large datasets that explore a wide range of control actions and physiological states. However, due to ethical concerns and patient safety risks, such clinical data is not easily accessible. Indeed, data collection would require deliberately inducing critical states, such as patients nearing the threshold of consciousness or in deep sedation. To overcome this limitation, we developed the Anesthesia Response Simulator, a comprehensive tool conceived to mimic the response of patients undergoing TIVA. The simulator is designed based on clinical practice, literature models, and medical experience and developed in Matlab and Python.[1] It provides an in-simulation environment in which RL agents can explore different policies and learn optimal behaviors that adapt to inter- and intra-patient variability. AReS focuses on three components of anesthesia: hypnosis, hemodynamics, and akinesia. Figure 5 provides a schematic overview of the simulator. This tool consists of five main elements (from left to right): 1) the **input** unit to indicate the type and the amounts of drugs administered, 2) **Pharmacokinetic** models to calculate plasma concentrations from the drug infusion rates, 3) **Pharmacodynamic** models to simulate the patient's physiological state, 4)

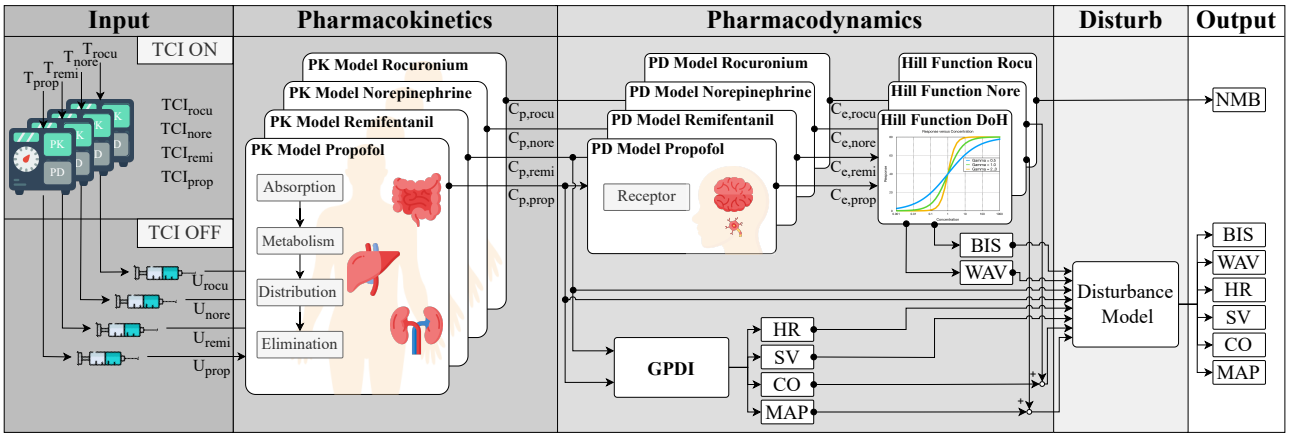


Figure 5: Schematic View of the Simulator.

Table 2: AReS Input variables and the corresponding measurement units.

Drug Category	Active Ingredient	TCI mode OFF	TCI mode ON
Anesthetic	Propofol	$u_{\text{prop}}(t)$ [mg/s]	$T_{\text{prop}}(t)$ [µg/mL]
Analgesic	Remifentanil	$u_{\text{remi}}(t)$ [µg/s]	$T_{\text{remi}}(t)$ [ng/mL]
Vasoconstrictor	Norepinephrine	$u_{\text{nore}}(t)$ [µg/s]	$T_{\text{nore}}(t)$ [ng/mL]
Neuromuscular blocking	Rocuronium	$u_{\text{rocu}}(t)$ [mg/s]	$T_{\text{rocu}}(t)$ [µg/mL]

the **disturbance** model to emulate the impact of surgical stimuli and blood loss on the patient response, and 5) the **output** unit to access the prediction of physiological states.

The simulator accounts for four drugs, namely propofol (anesthetic), remifentanil (analgesic), norepinephrine (vasoconstrictor), and rocuronium (neuromuscular blocking), described in Table 2. The input unit includes two modes where TCI can be either on or off. As shown in Table 2, when the TCI mode is off, the input variables are the infusion rates of the four drugs, $u_{\text{drug}}(t)$. In contrast, when the TCI mode is on, the input variables are the target concentration (either plasma or effect-site) of these drugs $T_{\text{drug}}(t)$. The inputs can be either fully predefined before the start of the simulation or decided step by step during the simulation, allowing the use of closed-loop control in the simulator.

As stated in Section 2, propofol, remifentanil, and rocuronium have been mentioned in the literature for their fast onset and recovery times [39, 66]. On the other hand, the administration of norepinephrine is rapidly becoming a standard for treating hypotension in the perioperative setting because of the reduced likelihood of organ damage [40]. The novel PK-PD models developed for norepinephrine and rocuronium in [40] and [21], respectively, provide a physiological representation between these drugs and commonly measured variables during anesthesia.

The simulator outputs are the anesthesia-related measurable variables detailed in Table 3. Specifically, DoH is measured by two indices: WAV_{cns} and BIS. The user decides which indexes are relevant for their purposes based on the application. The four variables MAP, CO, HR, and SV are the main variables to monitor the hemodynamic state. In AReS, the probability of observing the numbers of PTC is simulated for measuring profound NMB [21].

To initialize the simulation, in addition to selecting the TCI mode (ON or OFF), other parameters need to be configured. For propofol and remifentanil, the user should decide between several PK-PD models available. In the case of propofol, the PK-PD models are either population-based or subject-based. In the former case, the user only needs to set the patient's height, weight, age, and sex, while the other parameters are estimated from a base population. Conversely, in the latter case, the user must specify all the parameters described in detail in Section 5.2, as they are specific to a single patient. Moreover, one should set a disturbance profile, specifying quantities like the duration and intensity of the stimulus. More details about the parameters characterizing a disturbance profile will be provided in Section 5.4. The following sections give more information on PK models, PD models, and TCI systems.

[1]The package implementing AReS is available at GitHub `https://github.com/ManuMerlo/AReS-Anesthesia-Response-Simulator`.

Table 3: AReS Output variables and the corresponding measurement units.

Output	Description	Measurement Units
WAV_{cns}	NeuroWave index for DoH	percentage
BIS	Bispectral index for DoH	percentage
MAP	Mean arterial blood pressure	mmHg
CO	Cardiac output	L/min
HR	Heart rate	bpm
SV	Stroke volume	mL
NMB	Neuromuscular blockade	percentage

5.1. Pharmacokinetics models in AReS

Several studies have identified the relationships between patients' demographic parameters and flow rates, developing numerous population-based PK models [83]. In AReS, we implement specific PK models for each drug.

The three-compartment PK models developed by Schnider et al. [69] and Eleveld et al. [24] are integrated into the simulator for propofol. While the former model has been one of the most commonly used for TCI systems, the latter one is more recent and can be applied to a wider range of populations. Both models incorporate the influence of subject-specific covariates, such as age, weight, height, and sex, on propofol pharmacokinetic response. However, the Schnider model uses LBM, whereas the Eleveld model uses BMI. Moreover, the Eleveld model uses an allometric scaling approach and a maturation model to account for physiological differences, making it suitable for patients of various ages and BMI categories.

The remifentanil PK models employed in the simulator are the Minto [54] and the Eleveld [25] three-compartment PK models. Minto is a precedent and widely-used PK model that incorporates the influence of age and LBM on drug distribution and elimination. Minto et al. [54] focus on using and optimizing existing models for healthy adults, limiting the model's generalization to other demographics such as elderlies and children. In contrast, the Eleveld model incorporates data including adults and children and uses allometric scaling to account for the effect of body size on drug kinetics. Like for propofol, the Eleveld model considers age, weight, gender, and BMI of the subject to present physiological differences across different demographic groups. As a result, Eleveld et al. [25] find that the Eleveld model's performance exceeded existing models, such as Minto, especially for children and elderly people.

To address the challenge of intraoperative hypotension and its associated complications, we model low-dose norepinephrine administration during general anesthesia using the work by Joachim et al. [40] that proposes a two-compartment PK model with a depot.

Finally, the three-compartment model designed by De Haes et al. [23] is used to characterize the pharmacokinetics of rocuronium in both myasthenic patients and matched controls. Even in this case, the value of the coefficients for the PK model parameters depends on the patient's characteristics, i.e., they are proportional to their weight.

5.2. Pharmacodynamics models in AReS

As explained in Section 2, a PD model represents the mathematical relationship between the plasma concentration $C_p(t)$ and the physiological output $E(t)$, sometimes through an intermediate variable, the effect-site concentration $C_e(t)$. The PD models adopted for BIS, WAV_{cns} and NMB [21] comprise a linear component, which describes the relationship between $C_p(t)$ and $C_e(t)$, and a non-linear component which characterizes the relationship between $C_e(t)$ and $E(t)$. In contrast, the PD models we employed for the relationship between plasma concentration of propofol, remifentanil, and norepinephrine and hemodynamic variables directly relate $C_p(t)$ to $E(t)$. Furthermore, we extended the PD model presented by Su et al. [78] to jointly describe HR, SV, CO, and MAP, and include the norepinephrine effect on CO, and MAP.

The effect-site concentration of propofol, remifentanil, and rocuronium is computed using a linear model with a time delay as follows:

$$C_e(t+1) = k_{e0}(C_p(t-t_d) - C_e(t)), \tag{5}$$

where $C_e(t)$ is the effect-site concentration, k_{e0} is the drug's flow rate from plasma to effect-site compartments, and $t_d \geq 0$ is the time the drug takes to reach the effect-site compartment. In this simulator, time delay t_d reported in [38] is used for propofol effect-site concentration, while for remifentanil and rocuronium, t_d is

Table 4: The dataset used for identifying the relationship between $C_{\text{p,prop}}(t)$ and DoH for 44 subjects: subject demographics (upper part), PD model parameters identified in [38] (lower part). We report mean values and min-max ranges in brackets.

Demographics	Age	[yr]	55 [22-79]	
	Height	[cm]	173 151-190]	
	Weight	[kg]	87 [47-131]	
	Female	(male)	16 (28)	
			Eleveld PK	Schnider PK
PD Parameters	k_{e0}	[min^{-1}]	1.01 [0.145-9.46]	0.461 [0.108-2.37]
	T_d	[sec]	49 [0-150]	50 [0-150]
	EC_{50}	[µg/mL]	5.33 [2.30-8.32]	4.78 [2.17-9.10]
	γ		1.55 [1.15-1.78]	1.56 [1.13-1.80]
	E_0		91.5 [85.4-93.9]	91.5 [85.4-93.9]

considered zero.
In the case of linear or no interaction between drugs, the non-linear part of PD models is defined as the fractional sigmoid function or Hill function described in Equation (4). The parameter $C(t)$ is the effect-site concentration of drugs when calculating the DoH and NMB effect and the plasma concentration when computing the hemodynamic effect. In the case of surface interaction, we employed a variation of Equation 4, which is explained in detail later in this section.

5.2.1 Depth of Hypnosis

The PD model for calculating the DoH effect describes the relationship between the plasma concentrations of propofol and remifentanil and DoH monitoring signals (i.e., BIS and WAV_{cns} indices). Several PD models exist that describe the relationship between each drug and DoH. In particular, the studies we employed to model the Pharmacokinetics of remifentanil, [54] and [25], also include the PD models describing the relationship between remifentanil and BIS values. Similarly, the study in [24] includes the PD model for propofol. However, to account for inter-individual variability beyond population, another study in [38] is considered in this simulator that provides two sets of 44 subject-specific PD model parameters for computing the effect of propofol on WAV_{cns} index. The parameters of the PD model were identified in two distinct sets, depending on the PK model used to predict the plasma concentration of propofol. The first set of parameters was obtained by employing the PK model of Eleveld [24], while the second was derived using the PK model of Schnider [69].
AReS incorporates two interaction settings to calculate the DoH: *no interaction* and *surface interaction*. The *no interaction* setting ignores the effect of remifentanil and calculates DoH based solely on propofol PD models (e.g., Eleveld et al. [24]; Hosseinirad et al. [38]). Conversely, the *surface interaction* setting considers the effect-site concentrations of propofol and remifentanil in DoH calculations.
The surface interaction model developed in the work by Bouillon et al. [13] is implemented in AReS. Such a model provides the relationship between DoH and the effect-site concentration of propofol and remifentanil as follows:

$$\text{DoH}(t) = E_0 \left[1 - \frac{I^{\gamma}(t)}{1 + I^{\gamma}(t)}\right]. \tag{6}$$

The parameter $I(t)$ is the normalized ratio of the drug concentration with respect to the half-maximum effect drug concentrations EC_{50} and is defined as:

$$I(t) = \frac{U_{\text{remi}}(t) + U_{\text{prop}}(t)}{1 - \beta\theta(t) + \beta\theta(t)^2},$$

where, $U_{\text{prop}}(t) := C_{\text{e,prop}}(t)/EC_{50,\text{prop}}$ and $U_{\text{remi}}(t) := C_{\text{e,remi}}(t)/EC_{50,\text{remi}}$ are the normalized propofol and remifentanil effect-site concentration, respectively, and the parameter $\theta(t)$ is the potency of the drug mixture. More specifically, $\theta(t)$ is equal to 1 if only propofol is administered and 0 if only remifentanil is administered and is obtained formally as follows:

$$\theta(t) = \frac{U_{\text{prop}}(t)}{U_{\text{remi}}(t) + U_{\text{prop}}(t)}.$$

The parameter $\beta \geq 0$ has been modified such that the difference between the computed DoH with interaction and without the interaction is not more than 20% of the value of DoH if we disregard the interactions for all combinations of the propofol and remifentanil PK-PD models.
The sensors used to monitor BIS and WAV_{cns} from the scalp electroencephalogram signals have internal dynamics. For each monitor, either BIS or WAV_{cns}, Bibian et al. [10] identifies a Linear Time-Invariant (LTI) model to describe these internal dynamics. We implement these LTI models in AReS such that output signal BIS and WAV_{cns} are the multiplication of $\text{DoH}(t)$ computed in Equation (6) and the LTI model corresponding to each monitor.

5.2.2 Hemodynamics

As depicted in Figure 5, AReS considers the effects of three drugs, propofol, remifentanil, and norepinephrine, on hemodynamic variables. Specifically, the simulator employs the mechanism-based PD model, developed in [78] and based on the principles outlined in [75], to model the relationship between hemodynamic variables and propofol and remifentanil plasma concentrations. This framework incorporates the non-linear interaction between remifentanil and propofol using the general pharmacodynamic interaction (GPDI) model. The outputs of the model are Total Peripheral Resistance (TPR), HR, SV, CO, and MAP. In accordance with the principles in [75], MAP is considered the product of CO and TPR, and CO is the product of HR and SV, formally:

$$\text{MAP}(t) = \text{CO}(t) \cdot \text{TPR}(t) = \text{HR}(t) \cdot \text{SV}(t) \cdot \text{TPR}(t). \tag{7}$$

In the mechanism-based PD model, the inter-individual variability beyond demographics is incorporated by introducing stochasticity in the model parameters. In AReS, the user can choose whether the model parameters are their mean value or have variability caused by randomness. Detailed information regarding the parameters and non-linear interactions can be found in [78].
In AReS, the effect of norepinephrine on hemodynamic variables is also considered. The relationship between the plasma concentration $C_{\text{p, nore}}(t)$ and MAP identified in [40] is described using the Hill function in Equation (4). To the best of our knowledge, a PD model directly relating norepinephrine plasma concentration to CO has not been previously reported in the literature. However, [56] demonstrated that increasing the norepinephrine infusion rate from 0.2 to 0.4 µg/kg/min resulted in an 11% increase in CO. Therefore, in this simulator, the parameters of the Hill function correlating norepinephrine blood concentration to CO effect, EC_{50} and γ in Equation (4), are chosen such that an 11% increase in CO is achieved.
As far as we know, no existing study models the combined non-linear interactions of propofol, remifentanil, and norepinephrine on hemodynamic variables. Therefore, we propose a linear interaction between the effect of norepinephrine and the effects computed by the mechanism-based PD model for propofol and remifentanil, as follows:

$$\text{MAP}(t) = \text{MAP}_{\text{remi + prop}}(t) + \text{MAP}_{\text{nore}}(t), \tag{8}$$

$$\text{CO}(t) = \text{CO}_{\text{remi + prop}}(t) + \text{CO}_{\text{nore}}(t), \tag{9}$$

where $\text{MAP}_{\text{remi + prop}}(t)$ and $\text{CO}_{\text{remi + prop}}(t)$ are respectively the mean arterial pressure and cardiac output computed by the mechanism-based PD model in [78], $\text{MAP}_{\text{nore}}(t)$ is computed using PD model in [40], and $\text{CO}_{\text{nore}}(t)$ is computed using the model defined based on [56]. Both [56] and [42] discussed that norepinephrine does not have a significant effect on HR. Considering no effect on HR, the effect of all three drugs on SV is computed as follows:

$$\text{SV}(t) = \frac{\text{CO}(t)}{\text{HR}_{\text{remi + prop}}(t)}. \tag{10}$$

5.2.3 Neuromoscular Blockade

As illustrated in Figure 5, the NMB is decoupled from other outputs of the simulator, and rocuronium is the only drug that affects it. The parameters of the PD model specified in Equation (5) to calculate the concentration of rocuronium at the effect site from the plasma concentration are provided in [23]. In contrast, the non-linear PD model that provides the relationship between the concentration of rocuronium at the effect site and PTC is the one described in [21]. In the latter study, four categories are defined as the representation of an increase in NMB: $m = 0$ if PTC ≥ 6 (moderate), $m = 1$ if PTC $\in \{4, 5\}$ (deep), $m = 2$ if PTC $\in \{2, 3\}$ (profound), and $m = 3$ if PTC $\in \{0, 1\}$ (very profound). The PD model output is the vector $(p_0(t), \ldots, p_3(t))$ where $p_m(t)$ describes the probability of observing category m at time t.

5.3. Target controlled infusion

If users want to use the TCI in the simulator, they must configure the system as it would be in reality. The TCI can control the plasma or the effect-site concentration for propofol, remifentanil, and rocuronium. For norepinephrine, only the plasma-targeted TCI is available. Users can select the internal PK and PD models from those detailed in Sections 5.1 and 5.2, respectively. The plasma-targeted TCI uses the control algorithm described by Bailey and Shafer [9]. Conversely, the control strategy implemented in the effect-site-targeted TCI is designed by Shafer and Gregg [71]. To ensure safety, the maximum plasma concentration should be lower than the limit that causes intoxication. Moreover, the TCI pump shown in Figure 3 has a maximum drug delivery capacity. For these reasons, we incorporate the approach by Van Poucke et al. [82] in the TCI algorithm to restrict the plasma concentrations and limit the drug's infusion rate based on the capacity of infusion pumps. While the simulator allows users to set these limits as desired, we rely on medical studies to provide guidelines on configuring them and ensuring safety. Specifically, AReS uses 10 µg/mL and 2 mg/s as default limits for plasma concentration and infusion rate of propofol, respectively, according to [81]. Considering the capacity limits of the infusion pump in [81] and the drip concentration of remifentanil, norepinephrine, and rocuronium, their default infusion rate limits are 0.5 µg/s, 0.5 µg/s, and 0.1 mg/s, respectively. To our knowledge, there are no established safety limits for remifentanil, norepinephrine, and rocuronium plasma concentrations during anesthesia. Based on peak remifentanil concentrations observed using TCI as reported by Eleveld et al. [24], we set a default plasma concentration limit of 10 ng/mL for remifentanil. For norepinephrine and rocuronium, we empirically determined default limits by examining various concentration levels and selecting values that prevented unreasonable increases in MAP and NMB during the induction phase. These values are 15 ng/mL for norepinephrine and 10 µg/mL for rocuronium.

5.4. Disturbance Model

One of the primary objectives of the simulator is to provide an accurate representation of surgical stress. To achieve this, we incorporate two distinct types of disturbances: those deriving from external interventions, namely surgical stimuli, e.g., intubations and incisions, and those resulting from the patient's hypovolemic state, which may occur due to excessive fluid loss during the surgery. We model these disturbances differently based on their nature, i.e., disturbances originating from *surgical stimuli* are treated as additive, and those related to *hypovolemia* are represented as multiplicative terms.
More specifically, to model surgical stimuli, we examine the physiological mechanisms underlying the perturbations in the patient's state caused by external procedures. These actions are grouped into two categories: those causing tissue trauma and those that do not. Further analysis considers specific procedural types, such as intubation, incision, tissue manipulation, and suturing. Based on this context, we derive the corresponding response curves. Concerning hypovolemia, we adopted a multiplicative disturbances model approach supported by the clinical research.

5.4.1 Response to surgical stimuli

We define a surgical disturbance as follows:

$$y(t) := y_{\text{no_stimulus}}(t) + \Delta_{y,\text{stimulus}} \cdot f_{\text{stimulus}}(t), \tag{11}$$

where $y_{\text{no_stimulus}}$ is any output variable, e.g., WAV_{cns} or HR in the absence of the stimulus, $\Delta_{y,\text{stimulus}}$ is the amplitude of the response of output y stemming from the stimulus, and $f_{\text{stimulus}}(t)$ characterizes the dynamic of the disturbance based on the type of the surgical procedure, i.e., intubation, skin manipulation, suturing, and skin incision.
Although nociception monitors use indexes such as Analgesia Nociception Index (ANI) and Nociception Level (NoL) to measure the analgesic state, in clinical practice, it is typically assessed using HR and MAP, as these parameters are easily interpretable and can be measured continuously and noninvasively [36]. While the common trends about the changes occurring to CO, SV, and TPR in response to nociceptive stimuli are described by Andreis and Singer [4], the effects of surgical stimuli on these variables are not properly documented due to operational limitations. The autonomic response to nociception leads to an increase in HR, subsequently raising CO. Although variations in SV may occur depending on the specific mechanisms involved in the stress response, these variations are generally considered negligible for practical purposes. Moreover, nociception induces peripheral vasoconstriction, which leads to an increase in TPR. The latter is commonly considered the primary cause of the observed rise in MAP. Therefore, to simulate the hemodynamic nociceptive response, we model the effect of surgical stimuli on DoH, HR, and TPR using Equation (11). The effects of these disturbances subsequently influence CO and MAP due to their dependence on HR and TPR, as shown in Equation (7).

5.4.2 The amplitude of surgical stimuli

In this simulator, the amplitude of surgical stimuli $\Delta_{y,\text{stimulus}}$, introduced in Equation (11), is formally defined as follows:

$$\Delta_{y,\text{stimulus}} = \theta_{y,\text{stimulus}}[\lambda + (1-\lambda)\cdot w], \tag{12}$$

where $\theta_{y,\text{stimulus}} > 0$ is the maximum effect the patient can experience in response to the stress caused by the procedure. Users set the latter based on the intensity of the surgical stimulus they wish to apply. Finally, $\lambda \in (0,1)$ is the minimum proportion of $\theta_{y,\text{stimulus}}$ corresponding to the smallest patient's response, and $w \in [0,1]$ is the realization of a random variable W that weighs the patient's response based on their plasma drug concentrations. The principles of the up-and-down method [58] are the inspiration for defining the distribution of W, which is commonly used to determine the effective drug dose required to achieve a specific clinical outcome. This method observes whether a patient responds to a stimulus and adjusts drug doses based on a binary (yes/no) response. Similarly, we determine the distribution of W based on a binary response derived from the probability of non-response value introduced by Mertens et al. [53] and computed as follows:

$$P_{\text{no_response}} = \frac{(\epsilon \cdot C_{\text{p,prop}}(t) \cdot C_{\text{p,remi}}(t))^{\gamma}}{1 + (\epsilon \cdot C_{\text{p,prop}(t)} \cdot C_{\text{p,remi}}(t))^{\gamma}}, \tag{13}$$

where $C_{\text{p,prop}}(t)$ and $C_{\text{p,remi}}(t)$ are the plasma concentrations of propofol and remifentanil, respectively, $\gamma > 0$ is a parameter that describes the steepness of the dose-response curve, and $\epsilon \in (0, +\infty]$ is the interaction parameter. The parameters γ and ϵ depend on the type of stimulus applied to the patient. To the best of our knowledge, the only available values for these parameters are provided by Mertens et al. [53] for intubation stimuli. With slight abuse, we will use such values to calculate the probability of the patient's non-response for all surgical procedures.
The distribution $f(W)$ of the random variable W is formally defined as follows:

$$f(W) := \begin{cases} 1 - P_{\text{response}} & \text{if } W = 0 \\ P_{\text{response}} \cdot \frac{2(1-P_{\text{response}})}{P^2_{\text{response}}} \cdot W & \text{if } 0 < W \leq P_{\text{response}} \\ P_{\text{response}} \cdot \frac{2P_{\text{response}}}{(1-P_{\text{response}})^2} \cdot (1-W) & \text{if } P_{\text{response}} < W \leq 1 \end{cases},$$

where $P_{\text{response}} := 1 - P_{\text{no_response}}$. An example showing the distribution for different values for P_{response} is provided in Figure 6. In practice, with probability $P_{\text{no_response}}$ the patient is not responding ($W = 0$). Otherwise, we draw W from a distribution that favors coefficients close to 1 when P_{response} is large and coefficients close to 0 otherwise. This is given by the property of the distribution that the cumulated probability that the random variable W is in the range $[0, P_{\text{response}}]$ is equal to $1 - P_{\text{response}}$, thus favoring the extraction of a small coefficient.

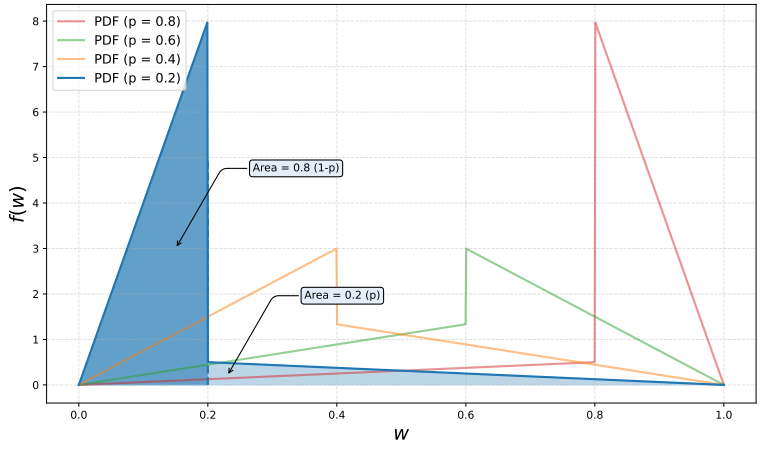


Figure 6: Probability Density Function (PDF) function for $W > 0$ with $P_{\text{response}} \in \{0.2, 0.4, 0.6, 0.8\}$.

Note that higher plasma concentrations increase the probability of no response $P_{\text{no_response}}$ in Equation 13. Consequently, the likelihood of a response P_{response} decreases. Using the distribution of W and simple mathematical manipulation, one can show that the expected value of W drops as P_{response} decreases. Thus, as plasma concentrations of propofol and remifentanil rise, the expected amplitude of surgical stimuli $\Delta_{\text{y, stimulus}}$ in Equation (12) reduces.

5.4.3 Characterization of the dynamics of stimuli's curves

We designed the model representing the dynamics of patients' reactions to surgical stimuli based on two key concepts: the physiological response of the sympathetic nervous system and the classification of applied stimuli. The former is based on releasing and eliminating catecholamines, including epinephrine, norepinephrine, and dopamine. These hormones and neurotransmitters play a central role in the body's *fight-or-flight* response to stress [61]. During painful events, epinephrine and norepinephrine interact with adrenergic receptors, triggering hemodynamic changes that help the body return to homeostasis [90]. As explained earlier, nociception increases HR, subsequently elevating CO, while also inducing peripheral vasoconstriction, which raises TPR and, in turn, MAP. Furthermore, research has demonstrated that the exogenous administration of epinephrine can raise the BIS level from 63 to 76 in lightly anesthetized patients, suggesting a significant arousal effect [5]. Since endogenous and exogenous epinephrine interacts with the same adrenergic receptor pathways, we expect that endogenous epinephrine produces a similar arousal effect. The main difference, as demonstrated by Schoffstall et al. [70], is that endogenous epinephrine production, even under conditions of maximal stress, does not reach the same plasma concentrations as those achieved by therapeutic administration. Consequently, while the fundamental mechanism remains the same, the magnitude of the response may be lower with endogenous epinephrine. Finally, Coleman et al. [19] demonstrate that BIS levels significantly increased during moderate and severe noxious stimuli compared to baseline BIS. However, the response magnitude was similar regardless of stimulus intensity.
While these physiological responses justify an increase in parameter values, categorizing stimuli based on the type of pain sensation is useful to understand the dynamics by which parameters return to baseline. Armstrong et al. [6] provide the characteristics of two classes of stimuli: the stimuli that do not cause tissue trauma and those that do. The first class of stimuli results in a rapid activation of the pain sensory pathway that ceases when the stimulus itself ceases. This class includes the laryngoscopy/intubation maneuver, the tractions exerted by the surgeon on tissues and nerve structures, and the placement of sutures at the end of surgery. These stimuli can be represented by curves that increase rapidly and decrease, albeit more slowly, at the end of the procedure. The second class is accompanied by the activation of an intense inflammatory reaction that prolongs the stress for at least 48 hours, ending with wound healing about 72 hours after trauma. Tissue trauma caused by surgical incision is the main stimulus that falls into this category. This stimulus can be represented by a rapid positive change in stress level followed by a minimal decreasing phase until the next 48 hours and a steep decreasing phase between 48 and 72 hours after the onset of trauma.

5.4.4 Body fluid balance and Circulatory volume

Maintaining an adequate intravascular volume throughout the surgical procedure is necessary to minimize risks of organ dysfunction, postoperative morbidity, and death [29]. The state where adequate intravascular volume is preserved is called *normovolemia*, whereas *hypovolemia* is defined as a state of reduced blood volume. Some of the most common causes of hypovolemia are fasting before surgery [14], blood loss associated with skin incisions and tissue manipulations, and evaporation of fluids from exposed wounds [45]. All these factors determine a reduction in the volume of blood filling the heart, resulting in a decrease in SV and CO. These scenarios are prevalent in operating rooms; therefore, we provide an approach to model hypovolemia using a multiplicative coefficient that modifies the value of a hemodynamic variable, such as MAP and CO, in the absence of hypovolemia, as follows:

$$y(t) = w_{\text{state},y} \cdot y_{\text{normovolemia}}(t). \tag{14}$$

Here, the coefficient $w_{\text{state},y}$ changes the output y according to the patient's state. If the patient is in a normovolemic state $w_{\text{normvolemia},y} = 1$ for all $y(t)$.
In our work, we employed findings by Skytioti et al. [73] to quantify $w_{\text{hypovolemia},y}$. For instance, the transition from a normovolemic to a hypovolemic state of a patient resulted in significant hemodynamic changes, i.e., SV decreased by 26%, CO decreased by 13.2%, MAP increased by 4%, and HR increased by 16%.
In AReS, we let the user directly change the state of the patient's blood volume by specifying the exact time during the simulation when they want the patient to switch to a hypovolemic state or return to a normal state.

6. Reinforcement Learning Framework

This section describes the simulator configuration that trains our RL agents. We also provide an overview of the algorithm employed, explaining the rationale behind our choice. Finally, we outline the hyperparameter configuration and the architectural decisions made during training.

6.1. Environment: Simulator Configuration

In the following, we describe the simulator configuration within the in-simulation environment we have developed. We set the simulator's TCI mode to off, allowing the agent to directly control the propofol and remifentanil infusion rates. Each episode lasts $T_{\text{sim}} = 5000$ seconds. The simulator sampling time is set to $T_s = 5$ seconds, meaning each action the agent selects is held constant for this duration before executing the next one. Formally, the action at time t is given by:

$$a(t) = a_k, \quad \text{for } t \in [kT_s, (k+1)T_s),$$

where a_k is the action taken at step k with $k \in \{0, 1, 2, \ldots, T_{\text{sim}}/T_s\}$. Consequently, the agent executes T_{sim}/T_s actions during each episode.[2]

The three agents are trained by simulating a different patient in each episode, with demographic data randomly generated. Specifically, the age and BMI category and the height are determined by sampling from the values summarized in Table 5.

Table 5: Age and BMI categories and Height with their corresponding ranges.

Parameter	**Category**	**Range**
Age [years]	Young	$[18, 40)$
	Middle-aged	$[40, 60)$
	Old	$[60, 80)$
BMI [kg/m^2]	Underweight	$[17, 18)$
	Normal	$[18, 25)$
	Overweight	$[25, 30)$
	Obese	$[30, 33)$
Height [cm]		$[140, 200)$

The final weight is derived from the selected BMI and height. This approach ensures that the population the agent is trained on includes a wide range of demographic data, enabling it to learn a policy conditioned by individual patient characteristics.

We use the Eleveld PK and PD models for propofol and remifentanil as they provide a more accurate representation across a diverse patient population. We deliberately avoid patient-specific approaches, as the dataset in the simulator contains individualized data for only 44 patients. Furthermore, since population-based PK-PD frameworks fix relationships with covariates such as age and weight (see Section 8.2), using the Eleveld models enables us to better understand whether the agent's policies align with the underlying dynamics of the simulator (see Section 8.4). Moreover, we consider the surface interaction between the two drugs.

6.2. Learning Algorithm

All three agents are trained with the same learning algorithm: Twin Delayed Deep Deterministic Policy Gradient (TD3) [28]. TD3 is an online model-free reinforcement learning algorithm that employs an actor-critic architecture, where the actor learns the policy and the critic evaluates it by estimating the action-value function $Q(s, a)$. TD3 is model-free, meaning it does not keep a model of the transition function P. This algorithm is based on the Deep Deterministic Policy Gradient (DDPG) [46] method, introducing modifications to address two key issues. The first is overestimation bias. Overestimation bias is a phenomenon occurring in some reinforcement learning algorithms in which maximization of a noisy value estimate induces a consistent overestimation [28]. To mitigate this problem, TD3 uses two critics instead of one, updating the policy based on the minimum estimated value. The second, training instability, is addressed by updating the policy less frequently than the critic, thus stabilizing learning and avoiding updates based on high-variance value estimates [28]. Finally, target policy smoothing is used to prevent the deterministic policy from overfitting to narrow peaks in the value estimate. This technique involves adding small Gaussian noise to the target actions.

Figure 7 summarizes the algorithm, which proceeds as follows. First, the critic networks Q_{θ_1}, Q_{θ_2}, and the actor network μ_ϕ are initialized with random parameters θ_1, θ_2, and ϕ. The target networks are then initialized with $\theta_1^{'} \leftarrow \theta_1$, $\theta_2^{'} \leftarrow \theta_2$, $\phi^{'} \leftarrow \phi$, together with the replay buffer $\mathcal{B}$ that stores and replays experiences collected from interactions of our agent(s) with the environment.

In each episode, the agent selects an action based on its **current policy**, to which a small Gaussian noise is added to encourage exploration (step 1):

$$a = \mu_\phi(s) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma).$$

[2]We use the PID in Section 7 to transition the patient from the induction phase to the maintenance phase.

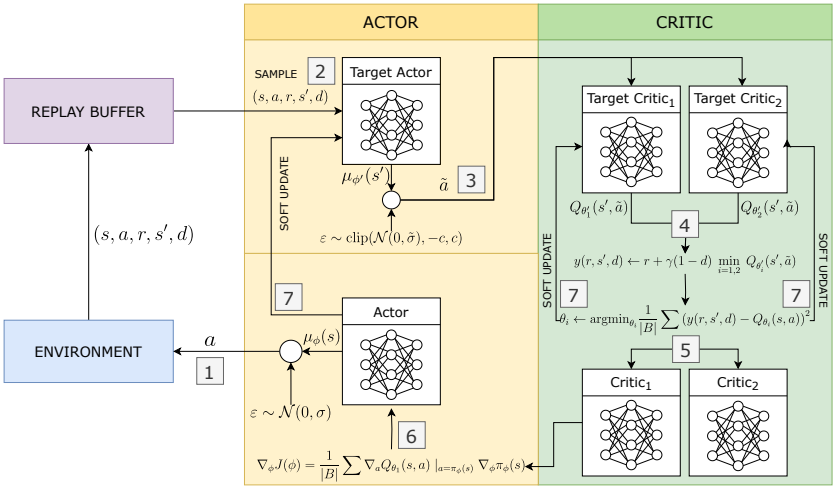


Figure 7: Overview of TD3 Algorithm.

After executing a, the agent observes the transition tuple (s, a, r, s', d), which is stored in the replay buffer $\mathcal{B}$. A mini-batch of transitions $B = \{(s, a, r, s, d)\}$ is then sampled from the buffer $\mathcal{B}$ (step 2). For each sample, the target value y for the critics is computed using the minimum of the two **target** Q-values (step 4):

$$y(r, s', d) \leftarrow r + \gamma(1 - d) \min_{i=1,2} Q_{\theta'_i}(s', \tilde{a}),$$

where d is the done signal, which is equal to 1 when the transition corresponds to the last step of an episode and 0 otherwise. $\tilde{a}$ is the action selected by the **target** actor given the state s', to which clipped Gaussian noise has been added (step 3):

$$\tilde{a} = \mu_{\phi'}(s') + \varepsilon, \quad \varepsilon \sim \text{clip}\left(\mathcal{N}(0, \tilde{\sigma}), -c, c\right).$$

Here, c is a clipping threshold that limits the magnitude of the noise added to the target action.
The critics are then updated by minimizing the loss function (step 5):

$$\theta_i \leftarrow \text{argmin}_{\theta_i} \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} \left(y(r, s', d) - Q_{\theta_i}(s, a)\right)^2, \quad i \in \{1, 2\}.$$

Every policy_delay update of the critics, the policy is updated with respect to Q_{θ_1} using the deterministic policy gradient (step 6), and target networks are updated using soft updates (step 7):

$$\nabla_\phi J(\phi) = \frac{1}{|B|} \sum_{s \in B} \nabla_a Q_{\theta_1}(s, a) \mid_{a=\mu_\phi(s)} \nabla_\phi \mu_\phi(s), \tag{15}$$

$$\theta_i^{'} \leftarrow \tau\theta_i + (1 - \tau)\theta_i^{'}, \quad i \in \{1, 2\}, \tag{16}$$

$$\phi^{'} \leftarrow \tau\phi + (1 - \tau)\phi^{'}. \tag{17}$$

We chose this algorithm because the three MDPs involve continuous action and state spaces. Moreover, in a healthcare context, a deterministic policy is preferable to a stochastic one. Stochastic policies produce a probability distribution over possible actions, introducing uncertainty into the decision-making process. In addition to being risky, this behavior may make the approach less acceptable to regulators accustomed to deterministic tools.

6.3. Hyperparameters and Architectural Design

The following provides an overview of the hyperparameter configuration and architectural choices used to train the three RL agents. They all share a common architecture where both the actor and critic networks are fully connected neural networks. Each network has two hidden layers, with 64 neurons in the first layer and 32 neurons in the second, using ReLU activation functions. In addition, the actor's output layer includes a Tanh activation function. The learning rate is fixed at 0.001 and, to encourage exploration, a zero-mean Gaussian noise is added to the actions, with a standard deviation (σ) that gradually decreases throughout each episode. The initial σ and decay rate values are agent and action-specific. The chosen optimizer is Adam [44]. A detailed summary of all hyperparameters used for training is provided in Table 6.

Table 6: Hyperparameters used to Train the RL Agents.

Hyperparameter	Agent 1	Agent 2	Agent 3
Learning Rate	0.001	0.001	0.001
Batch Size	256	256	256
Gamma	0.99	0.99	0.99
Tau	0.005	0.005	0.005
Optimizer	Adam	Adam	Adam
Replay Buffer Size	1M	1M	1M
Action Noise Propofol (σ)	0.1	0.1	0.1
Action Noise Remifentanil (σ)		0.001	0.001
Action Noise Decay Rate	0.99	0.999	0.999
Training Steps	70000	60000	80000

The main difference between the agents is the feature extraction mechanism. Agent 1 employs a standard Multi-Layer Perceptron (MLP) policy, which uses a simple FlattenExtractor that takes the input tensor and flattens it into a 1D vector. On the other hand, Agents 2 and 3 include a custom feature extractor that takes advantage of an idea influenced by Feature-wise Linear Modulation (FiLM)[64] layers to adapt feature representations based on conditioning variables.
A FiLM layer is formally defined as a feature-wise affine transformation that conditions an intermediate representation based on some external information. Specifically, if F is an intermediate feature map and x is a conditioning input, a FiLM layer modulates F through the affine transformation:

$$\mathrm{FiLM}(F \mid x) = \gamma\, F + \beta.$$

The scale and bias parameters are defined as:

$$\gamma = f(x) \quad \text{and} \quad \beta = g(x),$$

where f and g are learnable functions (often implemented as neural networks).
To incorporate this technique while maintaining compatibility with Stable-Baselines3 (SB3), we propose an adaptation of FiLM that conditions action selection based on demographic data without modifying the policy architecture. Specifically, instead of altering the policy network to accept an external conditioning input, we add a custom feature extractor whose architecture is shown in Figure 8 for Agent 2.
Given the state vector $x = [\text{Age, Height, Weight, Gender, WAV}_{\text{error}}]$, we first compute an expansion of the demographic data [Age, Height, Weight, Gender] using a neural network. The result is the conditioning context:

$$x_{\text{context}} = \mathrm{NN}_{\xi}(x_1, x_2, x_3, x_4),$$

where x_1 is the age, x_2 is the height, x_3 is the weight, and x_4 is the gender. NN_{ξ} represents a neural network with learnable parameters ξ.
A second neural network processes this conditioning input to generate the scale and bias terms:

$$\gamma, \beta = \mathrm{NN}_{\zeta}(x_{\text{context}}),$$

where NN_{ζ} is a learnable neural network parameterized by ζ. We then apply the FiLM operation only to the $\mathrm{WAV}_{\text{cns}}$ dimension:

$$\widehat{\mathrm{WAV}}_{\text{error}} = \gamma \cdot \mathrm{WAV}_{\text{error}} + \beta.$$

Unlike the FiLM approach, our version does not apply the affine transformation to the entire input. Finally, we concatenate the raw $\mathrm{WAV}_{\text{error}}$, the conditioned $\mathrm{WAV}_{\text{error}}$, and the conditioning data and process it with a subsequent neural network to integrate the information for action selection.
Figure 9 represents the feature extractor for Agent 3. Its architecture follows a similar structure to that of the feature extractor just described but introduces an additional conditioning step. Instead of a single scale-bias pair, it computes two sets of FiLM parameters to account for the presence of the $\mathrm{MAP}_{\text{error}}$ value:

$$\widehat{\mathrm{WAV}}_{\text{error}} = \gamma_1 \cdot \mathrm{WAV}_{\text{error}} + \beta_1,$$

$$\widehat{\mathrm{MAP}}_{\text{error}} = \gamma_2 \cdot \mathrm{MAP}_{\text{error}} + \beta_2.$$

The final 16 features are obtained by processing the raw $\mathrm{MAP}_{\text{error}}$, the raw $\mathrm{WAV}_{\text{error}}$, $\widehat{\mathrm{WAV}}_{\text{error}}$, $\widehat{\mathrm{MAP}}_{\text{error}}$, and the demographic data with the last combination neural network.

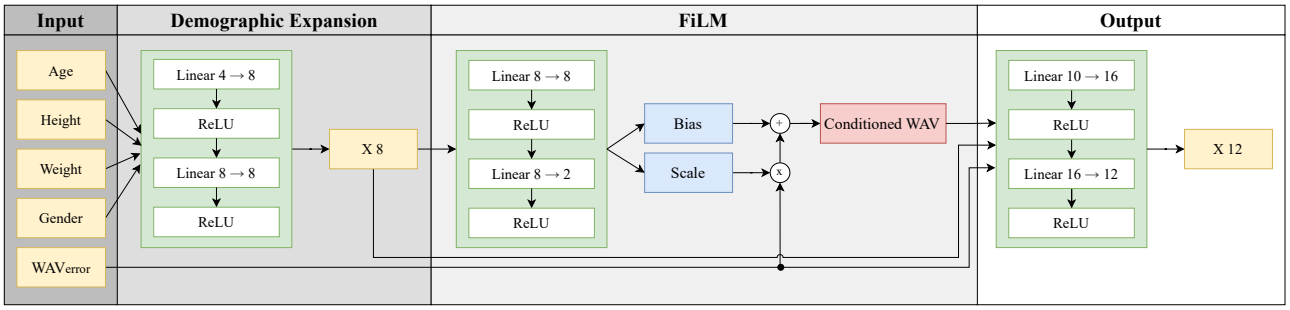


Figure 8: Feature Extractor Agent 2.

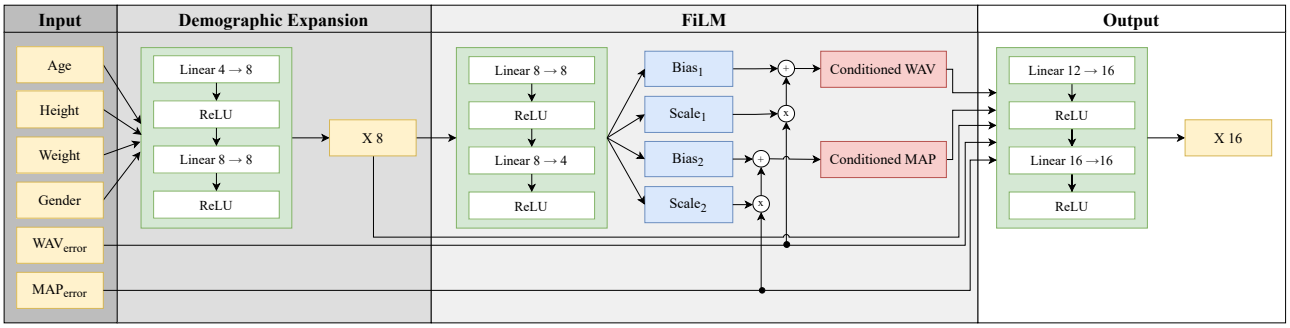


Figure 9: Feature Extractor Agent 3.

7. Traditional Control Approach: PID

To evaluate the performance of our RL agents, we establish a baseline using a Proportional-Integral-Derivative controller. In this work, we implement a PID controller inspired by [52] that regulates $\mathrm{WAV}_{\mathrm{cns}}$ during the induction and maintenance phases by using a fixed propofol-to-remifentanil ratio of 2 ($u_{\mathrm{remi}} = u_{\mathrm{prop}}/2$). As for the RL agents, external stimuli and intra-operative complications are disregarded. A schematic representation of the closed-loop system is shown in Figure 10.

A PID controller adjusts the control input based on the error between the desired and actual output. The control law is defined as:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)\, d\tau + K_d \frac{d}{dt} e(t), \tag{18}$$

where $e(t)$ represents the control error, and K_p, K_i, K_d are the proportional, integral, and derivative gains, respectively. To optimize the PID parameters, we employ a genetic algorithm [37] that iteratively adjusts the control gains. As shown in Figure 11, the process begins with an initial population of candidate PID parameter sets that are randomly generated. The performance of each candidate set is evaluated using the Integral of Absolute Error (IAE) as a metric. The objective is to minimize the total IAE across simulations of in-silico patients, whose demographic data are taken from 44 individuals [38]. The simulator is configured in the same way as it was for training the three RL agents (see Section 6.1).

During each simulation, the PID controller receives the current $\mathrm{WAV}_{\mathrm{cns}}$ value and determines the appropriate infusion rate for propofol. The absolute error is computed as:

$$e_{\mathrm{abs}}(t) = \left| 0.5 - \frac{\mathrm{WAV}(t)}{100} \right|.$$

The IAE is then accumulated over time:

$$IAE = \sum_{t=0}^{t_{\mathrm{sim}}/t_s} e_{\mathrm{abs}}(t) \cdot t_s,$$

where $t_{\mathrm{sim}} = 30$ min and $t_s = 5$ sec. Finally, the total fitness score is obtained by summing the IAE values across all patient simulations:

$$F_{\mathrm{total}} = \sum_{j=1}^{44} IAE_j.$$

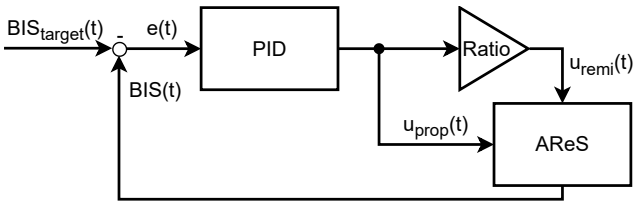


Figure 10: Schematic representation of the closed-loop control system using the PID controller.

The fitness scores are used to select the best-performing candidates. Specifically, the two best parameter sets are chosen using the roulette wheel selection. The probability that a parameter set i will be chosen as a candidate to create the next population is given by:

$$\text{prob}_i = 1 - \left(\frac{F_{\text{total},i}}{\sum_i F_{\text{total},i}} \right), \quad i \in \{1, \ldots, N\}.$$

The selected candidates, also referred to as *parents*, undergo crossover and mutation to generate a new population. Specifically, the crossover creates a *child* set whose parameters are the average of the parents' parameters with a probability of 0.8, otherwise, it returns the first parent. All children generated in this step then undergo the mutation process. For each parameter, with a probability of 0.2, a small value is added based on the magnitude of the parameter. This allows the algorithm to explore the parameter space over successive generations. The optimized PID parameters obtained via genetic algorithm are $K_p = 1.187$, $K_i = 0.001$ and $K_d = 0.204$.

8. Evaluation

In this section, we validate AReS as a tool to be used in the design of a decision-making system. Its performance is examined and evaluated to verify that the predicted output variables in response to different inputs are comparable with clinical data and analytically compatible with observations presented in previous studies. To this end, we compare the simulator's response with real data and evaluate its sensitivity to different model configurations. Furthermore, we assess the performance of the trained agents and analyze the behavior of those exploiting patients' demographics using explainable Machine Learning (ML). First, we assess their effectiveness based on key performance metrics for closed-loop control systems. Then, we apply explainability methods to gain insights into their decision-making processes and understand the factors influencing their actions.

8.1. AReS Performance Evaluation

Given the same infusion rate values, we assess whether the simulator outputs are compatible with actual signal measurements during anesthesia. The experimental process consists of 3 steps: 1) data taken from [89] and containing both WAV_{cns} and hemodynamic variables are selected, 2) the simulator is configured so that its setting is similar to the conditions under which the signals are measured, and 3) predicted and measured signals are compared by visual inspection and using the Median Absolute Error (MedAE) as a performance index.
The comparison dataset includes subjects with an average age of 65.9 years in the range of 54 to 76, an average height of 176.8 cm ($163 - 188$ cm), and an average weight of 86.0 kg ($66 - 108$ kg). Most subjects are male (9 males, 1 female). The reported variables are propofol infusion rate, remifentanil infusion rate, WAV_{cns} index, MAP, CO, HR, and SV. This dataset is a subset of the one presented by West et al. [89], which includes a partial recording of stats and WAV_{cns} signals of 150 patients during anesthesia. Among the 150 patients, we selected 10 for whom the records of all hemodynamic variables used in the simulator are available.[3]
While the length of recorded signals varies across subjects, all individuals have at least 30 minutes of data recorded after the start of propofol infusion. This duration encompasses both the anesthesia induction and maintenance phases, allowing for a comprehensive analysis of the simulated response. Therefore, we choose a 30-minute simulation period. In [89], norepinephrine was not administered, and the NMB monitoring is different than the PTC modeled in this simulator. Therefore, the effects of norepinephrine and rocuronium are excluded in this experiment. Before comparing the signals, the measured output signals are denoised by removing outliers and applying appropriate filters.

[3]Note that this dataset of 10 subjects and the one mentioned in Table 4 are both a subset of the dataset in [89]. Therefore, there is an overlap between these two datasets, i.e., the measurement of propofol infusion rate and WAV_{cns} index for three subjects in the comparison dataset are used in [38] to identify the subject-specified PD models.

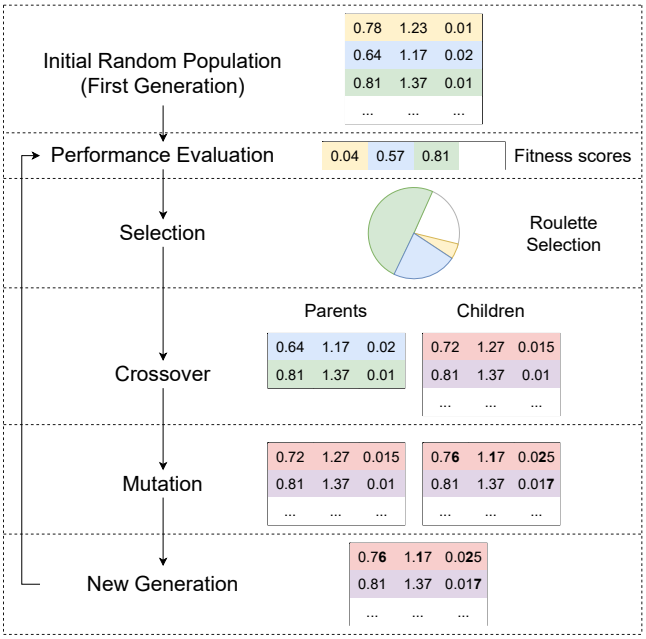


Figure 11: Genetic Algorithm Process.

Because the PK models of Eleveld [24, 25] are considered to represent a wider range of populations, including elderly and obese individuals, we chose these models to predict plasma concentrations of propofol and remifentanil. For propofol, we used the subject-specific PD models, while for remifentanil, we applied Eleveld's PD model. For subjects in the comparison dataset who are not included in the dataset described in Table 4, we selected the subject-specific PD model from those individuals in Table 4 with similar demographic characteristics. As a result, the demographic data used for the population-based PK-PD models came from the individuals in Table 4 chosen to represent the patients in the comparison dataset. We consider the surface interaction between propofol and remifentanil to compute the WAV_{cns}. In this implementation, we disregarded the inter-individual variability caused by randomness to have a fair comparison with the average population. To prevent differences in the initialization of variables from causing offsets between measured and simulated signals, the latter are initialized with the average values from the first 30 seconds of the corresponding measured signals. For each subject, the signal prediction error is quantified using the MedAE_y^i between the measured signals $y^i(t)$ of subject $i \in \{1, \ldots, 10\}$ and the predicted one $\bar{y}^i(t)$, formally:

$$\text{MedAE}_y^i = \text{median}\left\{|y^i(t) - \bar{y}^i(t)|\right\}_{t=0}^{t_{\text{sim}}}, \tag{19}$$

where t_{sim} is the time the simulation last and $y \in \{\text{WAV}_{\text{cns}}, \text{MAP}, \text{CO}, \text{HR}, \text{SV}\}$ is one of the outputs of the simulator. Given that the dimension of this error matches that of the evaluated parameter, it provides a meaningful value for comparison with the thresholds found in the literature. Additionally, considering the accuracy and limitations of clinical monitors, this error metric aligns with practical evaluation standards in anesthesia monitoring.

Results Figure 12 provides the predictions computed by AReS and compares them with the measured signals. Figures 12a and 12b provide the propofol and remifentanil infusion rates administered to the 10 subjects in the dataset, respectively. The infusion rates fluctuate since the dataset from [89] implements a closed-loop control of hypnosis using feedback from noisy WAV_{cns}. The plasma concentrations of propofol and remifentanil are simulated using the recorded infusion rates and are used to reproduce the output variables. Figures from 12c to 12g depict the results of simulated $\bar{y}(t)$ and measured output $y(t)$. Finally, we provide the values of MedAE for the five output signals and ten subjects in Table 7.

Discussion The median values of simulated hemodynamic variables in Figures 12c to 12f follow the median values of measured hemodynamic signals. As shown by comparing the shaded areas, at $t = 0$, the range of simulated hemodynamic variables is similar to the actual ones due to the initialization approach but decreases throughout the simulation. Thus, the simulated responses of each subject approach the population's median at the end of the simulation, suggesting that the demographical variability introduced by the simulator slightly underestimates the one present in real patients. Note that the inter-individual variability caused by randomness is disregarded in this experiment.

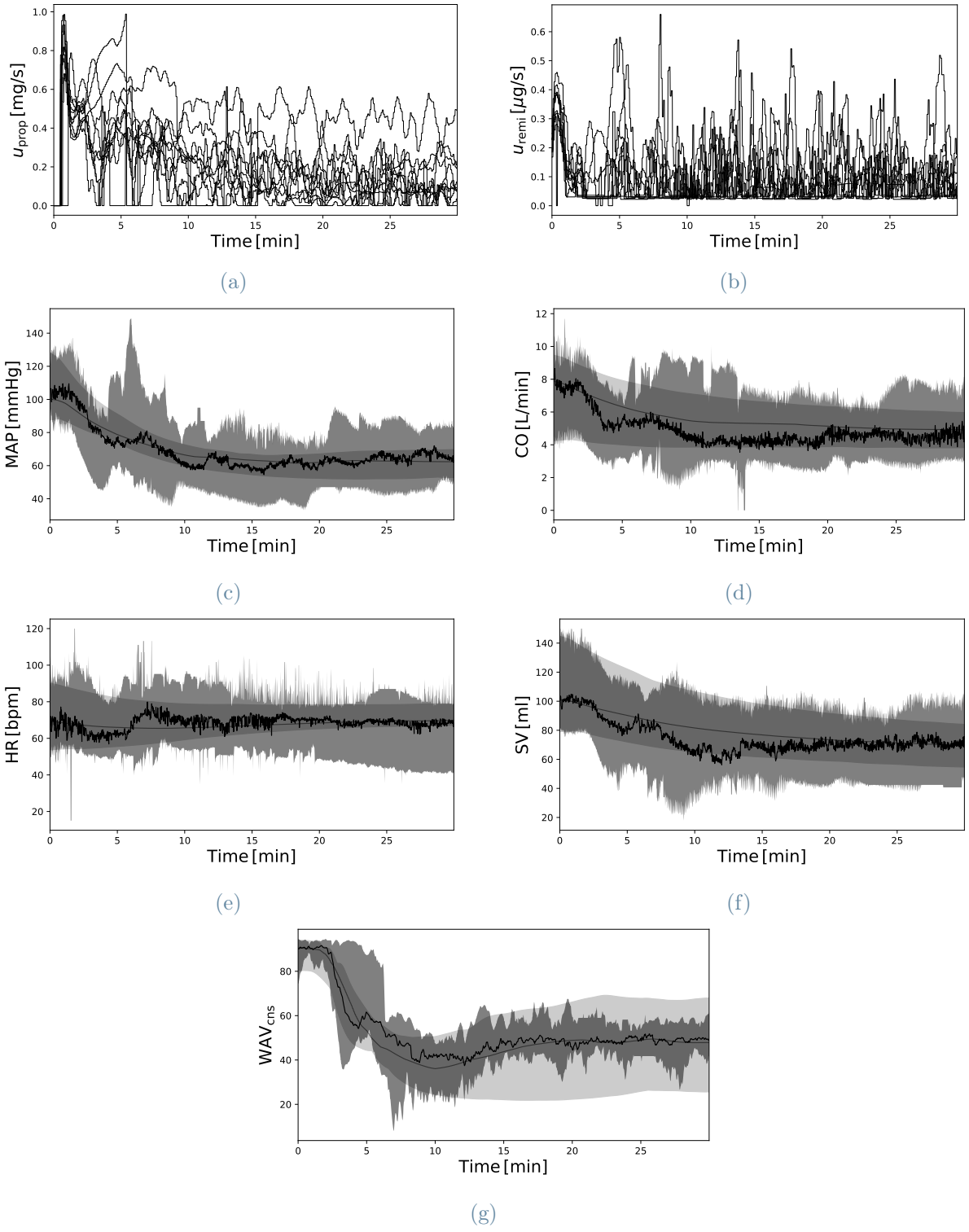


Figure 12: Figures (a) and (b) show the propofol and remifentanil infusion rates of the 10 subjects. In Figures (c) to (g), the lighter and darker black lines indicate the median of simulated outputs and actual denoised signals, respectively. The lighter and darker shaded areas show the range of simulated and actual signals, respectively.

The acceptable absolute error for measuring MAP is less than 10 mmHg [88], depending on the clinical context and the measurement method used. The MedAE values in Table 7 indicate that, except for subjects 5, 6, and 7, the predicted MAP signals are sufficiently close to the clinical ones. Zhou et al. [95] report that the maximum relative error in HR measurements, based on the sampling rate of the electrocardiogram, is 10%. If we consider 85 bpm as the mean value of HR, the maximum acceptable MedAE of HR is 8.5 bpm. Except for subjects 1 and 8, the MedAE of HR in Table 7 are less than this value, which indicates that the HR signals

Table 7: MedAE for the 10 Subjects.

	Subjects									
Output	1	2	3	4	5	6	7	8	9	10
WAV_{cns} [%]	4.41	6.17	18.56	20.6	5.59	7.53	6.26	4.77	20.0	5.75
MAP [mmHg]	9.02	8.71	8.29	3.28	16.08	11.23	18.48	8.19	7.33	9.39
CO [ng/mL]	1.42	0.751	1.03	1.30	0.608	0.905	1.04	1.20	0.621	0.644
HR [bpm]	9.39	2.66	4.01	5.98	3.51	7.45	7.87	9.02	3.18	7.71
SV [mL]	8.17	13.82	18.85	12.42	5.94	9.42	10.73	6.72	5.35	8.54

are comparable to clinical ones in most cases. In [22], the acceptable error percentage for measuring CO during anesthesia is generally considered to be 30% of the mean value or less. Considering the mean CO to be 6 L/min, the MedAEs of all subjects in Table 7 are less than the measurement error of 1.8 L/min. In [31], the 20% measurement error of SV mean value is considered acceptable, and in Table 7, assuming the mean value of SV is 80 mL, the MedAEs for SV is less than 16 mL for most subjects. These observations indicate that AReS's predictions of hemodynamic variables represent the trends observed in the average population response to propofol and remifentanil.
The WAV_{cns} median value of the simulated signals in Figure 12g also follows the median value of the measured signals. Compared to the hemodynamic variables, the inter-individual variability during the steady state is larger than the measured signals since the propofol subject-specific PD model [38] is used in the simulation, and these PD models cause inter-individual variability in dynamics during anesthesia. By inspecting the initial values of WAV_{cns} index in Figure 12g, we can observe that the variability in t_d described in Equation 5 as the delay in the response of the propofol subject-specific PD models is underestimated. This result is expected as the maximum PD models' delay considered by Hosseinirad et al. [38] is 3 minutes. However, as shown in Figure 12g, the PD models' delay for at least one of the subjects is more than 5 minutes. Moreover, in [72], less than five unit differences in DoH are considered clinically negligible. Except for subjects 3, 4, and 9, the others have MedAEs approximately 5 units, which can be considered negligible compared with the median signals magnitude.
This experiment demonstrates that, on average, across the population, the simulator response to propofol and remifentanil adequately represents the actual signals, making it suitable for designing an anesthesia decision-making system. However, due to inter-individual variability, achieving perfect predictions for every subject is unrealistic. This observation is supported by the results in Table 7, where some subjects exhibit MedAE values exceeding acceptable measurement errors. Therefore, while individual predictions may vary, the simulator provides a valuable tool to develop anesthesia decision-making systems at the population level.

8.2. AReS Sensitivity analysis

We design this experiment to examine the variability caused by the choices of the PK-PD model across the population. The primary goal is to demonstrate how the DoH varies when using different combinations of PK-PD models. The secondary purpose is to evaluate inter-individual variability in both DoH and hemodynamic variables within the same model across the population.
This experiment begins by setting the TCI mode on and the target effect-site concentrations for propofol and remifentanil to 2 µg/mL and 3.6 ng/mL, respectively, as experts usually suggest. The infusion rate and plasma concentration limits for TCI are set as the default values introduced in Section 5.3. Since only one model is available for norepinephrine and rocuronium, their administration would result in a uniform value shift of other variables. Therefore, we choose not to consider them. With fixed target concentrations of propofol and remifentanil, we analyze the six PK-PD model scenarios listed in Table 8. In this experiment, the PK-PD model for the TCI setting is set equal to the PK-PD model chosen in the corresponding scenario. In this way, the variability caused by the discrepancies between the model of the simulator and the controller is avoided. Finally, the inter-individual variability for each scenario is assessed by running experiments for each one of the individuals in the dataset described in Table 4.

Results Table 9 presents the infusion rate at steady state, the peak plasma concentrations of propofol and remifentanil, and the steady-state DoH, with the highest values for the peak concentrations and DoH highlighted in bold. For all scenarios, MAP reaches 71.7 ± 9.9 mmHg, CO stabilizes at 4.7 ± 0.5 L/min, HR at 67.8 ± 2.2 bpm, and SV at 70.0 ± 8.2 mL. Note that across different scenarios, steady-state hemodynamic variables remain unchanged. The reason for this is that MAP, CO, HR, and SV are directly influenced by

Table 8: Setting Options for the PK-PD Models of Propofol and Remifentanil, and the TCI Algorithm.

Scenario	Propofol PK	Propofol PD	Remifentanil PK	Remifentanil PD
1	Eleveld	Eleveld	Eleveld	Eleveld
2	Eleveld	Eleveld	Minto	Minto
3	Eleveld	subject-specific PD	Eleveld	Eleveld
4	Eleveld	subject-specific PD	Minto	Minto
5	Schnider	subject-specific PD	Eleveld	Eleveld
6	Schnider	subject-specific PD	Minto	Minto

Table 9: From left to right, steady-state infusion rates for propofol and remifentanil, peak plasma concentrations for propofol and remifentanil, and steady-state DoH. Data are presented as mean ± standard deviation (std).

	u_{prop} [mg/s]	u_{remi} [µg/s]	Peak $C_{\text{p,prop}}$ [µg/mL]	Peak $C_{\text{p,remi}}$ [ng/mL]	DoH percentage
Scenario 1	0.09 ± 0.01	0.19 ± 0.03	**8.88 ± 0.61**	4.93 ± 0.72	37.72 ± 2.30
Scenario 2	0.09 ± 0.01	0.15 ± 0.02	**8.88 ± 0.61**	**5.84 ± 0.66**	35.01 ± 5.28
Scenario 3	0.09 ± 0.01	0.19 ± 0.03	4.60 ± 1.63	4.93 ± 0.72	**49.58 ± 5.11**
Scenario 4	0.09 ± 0.01	0.15 ± 0.02	4.60 ± 1.63	**5.84 ± 0.66**	46.71 ± 5.40
Scenario 5	0.11 ± 0.01	0.19 ± 0.03	7.42 ± 2.48	4.93 ± 0.72	47.11 ± 5.68
Scenario 6	0.11 ± 0.01	0.15 ± 0.02	7.42 ± 2.48	**5.84 ± 0.66**	44.36 ± 6.49

the plasma concentrations of propofol and remifentanil, and are all calculated using the same mechanism-based PD model described in Section 5.2.2. Since all scenarios target the same concentration at the effect site, they ultimately reach the same steady-state plasma concentrations.
Figure 13 presents a boxplot representation of the simulation results for the DoH index across the scenarios. Scatter points denote steady-state values of DoH for each subject, and the lines connect the same individual across scenarios. The red line highlights the subject who experiences the largest change across scenarios. In each boxplot, the central blue line indicates the median, while the bottom and top edges of the box represent the 25-th and 75-th percentiles, respectively. The whiskers extend to the remaining data points that are not considered outliers.

Discussion We observe that across different scenarios mentioned in Table 8, when the propofol PK-PD model is the same (scenarios $1-2$, $3-4$, and $5-6$), the mean and variability in the peak of propofol plasma concentration are also the same. A similar trend occurs for the plasma concentration of remifentanil when the same PK-PD model of remifentanil is set. We note that the plasma peak depends on the PD model because the TCI operates in effect-site mode. Specifically, for propofol, the largest plasma concentration peak occurs with the Eleveld-Eleveld combination, followed by the Schnider-subject-specific PD combination, and finally, the Eleveld-subject-specific PD combination. As the peak of $C_{\text{p, prop}}$ for scenario 1 is larger than the one for scenario 2, we conclude that, on average, the subject-specific PD models have faster dynamics ($C_{\text{e, prop}}$ reaches the target faster) than the Eleveld PD model by referring to [71]. For remifentanil, the largest peak is achieved with the Minto-Minto combination, showing a difference of approximately 1 unit compared to the Eleveld-Eleveld combination.
Results in Table 9 show that when employing the same PK-PD model for propofol, DoH is consistently lower by about 3 units when using the Minto model instead of the Eleveld model for remifentanil, which is not a significative difference based on study in [72]. Note that $C_{\text{e, prop}}$ and $C_{\text{e, remi}}$ steady-state values match the target value for all scenarios. Therefore, according to Equation (4), the variation in the steady-state DoH across scenarios is due to the difference in the parameters EC_{50} and γ for propofol and remifentanil. Based on this, we compare steady-state DoH values for scenarios with the same remifentanil models (scenarios $1-3-5$ and $2-4-6$). It is worth noting that the nominal values of EC_{50} and γ parameters estimated by Eleveld's population-based PD model for propofol are 2.73 µg/mL and 1.89, respectively. Conversely, as shown in Table 4, the mean values of EC_{50} and γ for the subject-specific PD model are 5.33 µg/mL and 1.55 or 4.78 µg/mL and 1.56 µg/mL for the model estimated when using Eleveld or Schnider as the PK model, respectively. Since γ is not significantly different among the scenarios, the main difference between their steady-state DoH is due to the

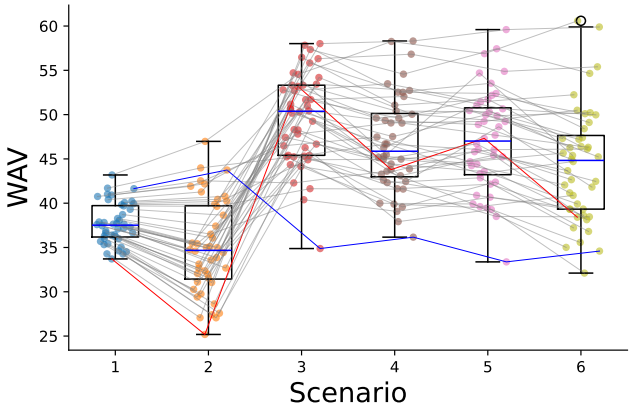


Figure 13: Steady-state DoH values for subjects across the scenarios in Table 8.

difference between the average EC_{50} of the employed PD models. According to Equation 4, larger EC_{50} results in fewer drops in DoH and higher steady-state values. Consequently, the lowest and highest DoH in Table 9 correspond to the propofol Eleveld-Eleveld and Elveld-subject-specific PK-PD models, respectively. This trend is also observed in Figure 13.

In this figure, comparing the variation in steady-state DoH of scenarios 1 and 2, we note that the former results in the least variability. This is partly because Eleveld's PD model for remifentanil estimates a parameter EC_{50} that is constant for the entire population. In contrast, the parameter EC_{50} in the Minto's PD model for remifentanil depends on patient demographic characteristics, such as the age [54], resulting in a larger dispersion of values. Observing the lines from scenario 1 to scenario 2, we see that they do not intersect, which allows us to deduce consistency between the scenarios. The reason is that the parameters remain consistent between individuals with similar demographic characteristics in the first two cases. In comparison, this assumption is no longer valid in the last four cases, where propofol PD parameters are estimated individually for each subject. For example, in these scenarios, it is now inappropriate to assume that the scatter points are ordered according to the age of the associated individual from oldest to youngest. In addition, the parameter EC_{50} of Eleveld's PD model for propofol, which depends on the age of subjects [24], remains in a range that for the 44 subjects is $[2.33, 3.35]$ µg/mL. The latter is smaller than the range for the subject-specific PD reported in Table 4. Therefore, the change in DoH due to propofol is smaller in scenario 1 than in scenarios 3 and 5. In Figure 13, the red line represents a 79-year-old man, one of the oldest individuals in the dataset. For this reason, in the first two scenarios, the DoH values are among the lowest. However, his estimated EC_{50} value of 6.40 µg/mL for scenarios 3 and 5 slightly exceeds the mean value of 5.5 µg/mL observed across all subjects. Consequently, the steady-state WAV_{cns} in these two scenarios is slightly elevated compared to the median. In contrast, the blue line represents a 32-year-old man, the third-youngest individual in the dataset. His estimated EC_{50} values for the last four scenarios are among the lowest, which places the corresponding DoH values at the whisker's limit. These results show that the inter-individual variability might not be all captured through demographic dependency.

8.3. Controllers Performance Evaluation

In this section, we evaluate the performance of the PID and RL-based controllers using AReS. The simulator is configured identically to its setup during the training process. However, instead of generating patients with randomly assigned demographic data, we use the data of 44 real individuals presented in [38]. Additionally, we repeat the evaluation by adding disturbances to mimic state perturbations caused by external stimuli during the surgical procedure to assess the robustness of the policies. Specifically, we reduce the simulation time to 30 minutes, introducing an incision at $t \in [2, 4]$, a tissue manipulation at $t \in [10, 20]$, and as a final procedure, a suture, at $t \in [22, 27]$. The maximum response amplitudes $\theta_{y,\text{stimulus}}$ (see Equation 12) for each stimulus and output y are presented in Table 10. As far as we know, there is little to no documentation on the amplitude of patient responses to surgical stimuli that are not heavily influenced by the depth of anesthesia. In light of this consideration, we take the maximum amplitude for MAP and HR responses during incision from [91] and the amplitude for the WAV_{cns} from [19]. Given that we have no reference for skin manipulation and suturing, we assume that skin manipulation is half the amplitude of the incision and that suturing is half the amplitude of skin manipulation. To ensure reliable comparison, we remove any randomness and set the w values (see Equation 12) to 1, representing the worst-case scenario in which every patient responds at the maximum level. We compare the performances using standard closed-loop infusion control metrics, as defined by Varvel et al.

Table 10: Maximum WAV_{cns}, HR, and MAP response amplitudes for each surgical procedure.

Surgical Procedure	$\theta_{\text{WAV}_{\text{cns}}}$	θ_{HR}	θ_{MAP}
Skin Incision	5	20	30
Tissue Manipulation	3	10	15
Suturing	2	5	7

[84], with the formula definitions and thresholds for the DoH index taken from [57]. We assume the thresholds for MAP are the same.
These metrics are based on the Performance Error (PE), formally expressed as:

$$\text{PE} = \frac{\text{Value} - \text{Target}}{\text{Target}} \cdot 100. \tag{20}$$

To assess control bias, we use the Median Performance Error (MDPE), which quantifies the central tendency of the PE for a single patient:

$$\text{MDPE}_i = \text{median}(\text{PE}_{ij}), \quad j \in \{1, \ldots, N\}, \tag{21}$$

where i represents a subject, and j indexes the PE measurements for that subject.
The Median Absolute Performance Error (MdAPE) measures the overall accuracy of the controller for a given subject:

$$\text{MdAPE}_i = \text{median}(|\text{PE}_{ij}|), \quad j \in \{1, \ldots, N\}. \tag{22}$$

We evaluate intra-subject variability using the Wobble metric, which captures fluctuations in PE relative to the MDPE:

$$\text{Wobble}_i = \text{median}(|\text{PE}_{ij} - \text{MDPE}_i|), \quad j \in \{1, \ldots, N\}. \tag{23}$$

Furthermore, we report the percentage of measurements in which the measured variable is observed to be within $\pm$ 5 Target. Finally, we present the Root Mean Square Error (RMSE).

Results Table 11 summarizes the performance of the four controllers in regulating the WAV_{cns} and MAP parameters, with the metrics calculated relative to target values of 50 and 70, respectively. Table 12 presents the same performance metrics for the same experiment but with disturbances introduced to simulate external stimuli. To facilitate interpretation, we highlight in bold the smallest errors across controllers and the highest percentage of time within the target ± 5 range. Additionally, Table 13 presents the total amounts of propofol and remifentanil administered by each controller in both scenarios. All values, including the performance metrics and drug amounts, are calculated for each individual and displayed as mean $\pm$ standard deviation. Figure 14 and Figure 15 visually display the WAV_{cns} and MAP trends, helping to better understand the performance metrics and their implications. In these figures, each line represents one of the 44 patients. The green areas represent safe regions for the variables, while the red ones mark risky scenarios. The red vertical lines indicate the onset of a disturbance.

Discussion First, we examine the case without disturbances. It can be observed that all performance metrics remain well below the threshold that defines the acceptable error margin. All the MDPE concerning the WAV_{cns} index in Table 11 are close to zero, showing that the four controllers achieve accurate sedation control near the target. Notably, in all cases, the MDPE is smaller than the relative MdAPE and slightly negative, indicating a minimal tendency to undershoot the WAV_{cns} target. Agent 1 and 2, which aim at controlling only the WAV_{cns} index, show slightly smaller MDPE than the PID. In contrast, Agent 3, which also aims to control hemodynamic stability, demonstrates a slightly higher, yet still small, MDPE error. When considering the MdAPE values for the PID, first, second, and third agents, which are 0.38, 0.1, 0.73, and 1.07, respectively, we observe that the predicted WAV_{cns} ranges from 49.90 to 50.10 for the first agent, 49.81 to 50.19 for the PID, 49.46 to 50.54 for the second agent, and 49.27 to 50.73 for the third agent. Additionally, the minimal differences in RMSE across these controllers (all less than 1) highlight that the performances are comparable.
The PID controller does not actively manage MAP, and its MDPE (-2.62%) corresponds to a slight tendency for the patient's MAP to remain below 70. Agent 1, which mirrors the PID strategy, shows a very similar bias (-2.59%). Agent 2, which controls both propofol and remifentanil to manage only the WAV_{cns} target, displays a smaller still negative bias for MAP (-1.39%). This suggests that minimizing remifentanil usage to

Table 11: Performance metrics for PID and RL Agents under normal conditions (83 min).

	Metric	PID	Agent 1	Agent 2	Agent 3	Threshold
WAV_{cns}	MDPE (%)	-0.24 ± 0.29	**-0.10 ± 0.17**	-0.20 ± 0.75	-0.87 ± 0.41	± 5.0
	MDAPE (%)	0.38 ± 0.16	**0.10 ± 0.17**	0.73 ± 0.29	1.07 ± 0.19	7.5
	Wobble (%)	0.20 ± 0.16	**0.03 ± 0.05**	0.09 ± 0.04	0.74 ± 0.42	5.0
	Controlled ±5 (%)	99.92 ± 0.51	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	≥ 80
	RMSE	0.51 ± 0.35	**0.16 ± 0.05**	0.37 ± 0.14	0.65 ± 0.07	5.0
MAP	MDPE (%)	-2.62 ± 3.58	-2.59 ± 3.59	-1.39 ± 3.59	**0.94 ± 4.25**	± 5.0
	MDAPE (%)	3.34 ± 3.14	3.13 ± 3.13	**2.72 ± 2.72**	3.70 ± 2.32	7.5
	Wobble (%)	**0.17 ± 0.07**	0.19 ± 0.07	0.29 ± 0.07	0.51 ± 0.06	5.0
	Controlled ±5 (%)	86.32 ± 33.98	86.36 ± 34.31	88.08 ± 31.75	**93.05 ± 21.00**	≥ 80
	RMSE	2.22 ± 2.22	2.24 ± 2.17	**2.02 ± 1.89**	2.60 ± 1.46	5.0

Table 12: Performance metrics for PID and RL Agents under disturbed conditions (30 min).

	Metric	PID	Agent 1	Agent 2	Agent 3	Threshold
WAV_{cns}	MDPE (%)	**-0.24 ± 1.86**	-0.26 ± 0.22	0.28 ± 0.52	-3.23 ± 0.42	± 5
	MDAPE (%)	2.13 ± 0.78	**1.03 ± 0.24**	1.51 ± 0.32	3.84 ± 0.41	7.5
	Wobble (%)	1.64 ± 0.25	**0.97 ± 0.28**	1.44 ± 0.34	2.43 ± 0.21	5.0
	Controlled ±5 (%)	98.62 ± 2.79	**100.00 ± 0.00**	**100.00 ± 0.00**	99.23 ± 1.93	≥ 80
	RMSE	1.62 ± 0.48	**1.01 ± 0.10**	1.25 ± 0.18	2.33 ± 0.25	5.0
MAP	MDPE (%)	37.87 ± 3.65	**37.74 ± 3.28**	39.00 ± 3.38	40.66 ± 3.59	± 5.0
	MDAPE (%)	37.87 ± 3.65	**37.76 ± 3.28**	39.00 ± 3.38	40.66 ± 3.59	7.5
	Wobble (%)	1.54 ± 0.21	**1.51 ± 0.17**	2.34 ± 0.25	3.43 ± 0.18	5.0
	Controlled ±5 (%)	9.70 ± 3.27	9.73 ± 3.30	9.73 ± 3.29	**9.75 ± 3.32**	≥ 80
	RMSE	**26.46 ± 2.30**	26.51 ± 2.20	27.12 ± 2.22	27.92 ± 2.27	5.0

Table 13: Drug administration for PID and RL Agents under normal and disturbed conditions.

Conditions	T_{sim}	Drug	PID	Agent 1	Agent 2	Agent 3
Normal	83 [min]	Propofol [mg]	448.79 ± 80.80	449.95 ± 81.86	477.93 ± 88.87	536.94 ± 122.08
		Remifentanil [μg]	224.39 ± 40.40	224.97 ± 40.93	217.89 ± 42.70	31.83 ± 16.89
Disturbed	30 [min]	Propofol [mg]	207.46 ± 37.02	202.97 ± 39.30	217.89 ± 42.70	266.61 ± 56.53
		Remifentanil [μg]	103.73 ± 18.51	101.48 ± 19.65	71.77 ± 16.82	5.02 ± 1.57

control sedation helps maintain slightly higher blood pressure. Finally, Agent 3, which actively targets both wav = 50 and MAP in the range of $[65, 75]$ mmHg, is the only controller that exhibits a slight positive MDPE (+0.94%), meaning it is more likely to exceed 70 mmHg rather than fall below it. From a clinical point of view, maintaining MAP slightly above 70 can be safer than risking lower pressures, particularly in patients susceptible to hypotension. The greatest difference of explicitly managing MAP can be seen in the "Controlled ±5" column. The PID controller and Agent 1 keep MAP within ±5 of 70 about 86% of the time, while Agent 2 slightly improves that percentage ($\sim 88\%$) and Agent 3 demonstrates the highest value ($\sim 93\%$). The "Wobble" metric, which captures variability around the target, remains low for all controllers and all indexes, implying no excessive oscillations. When comparing the WAV_{cns} and MAP MDPE errors for the third agent, we notice that the controller manages the first index better. However, this can be attributed to the fact that it focuses on maintaining MAP within the range of $[65, 75]$ rather than strictly equal to 70. In some sense, the way the agent has been trained in the third case prioritizes WAV_{cns} over MAP. This is also because, in clinical practice, other types of drugs (e.g., vasoconstrictors) and approaches such as fluid administration can be used to regulate MAP.

Table 13 shows the amount of drugs administered by the four controllers. Interestingly, the PID and Agent 1 usage of propofol and remifentanil are almost identical. Agent 2 administers more propofol and less remifentanil. By favoring propofol, Agent 2 helps preserve blood pressure somewhat better. This is in line with the behavior of the third Agent that uses the most propofol overall and the least remifentanil. Figure 16 and Figure 17 show that propofol has a greater effect on WAV_{cns} than MAP, while remifentanil influences both similarly. As a result, Agent 3 likely prioritizes propofol to maintain a WAV_{cns} of 50 while attempting to increase MAP.

However, while this strategy might be effective in normal conditions, it may not be optimal when considering disturbances. In this scenario (Table 12), all three agents manage to control the depth of hypnosis, keeping

performance errors within acceptable thresholds, but struggle to manage MAP.
This is expected for the PID controller and the first two RL agents, as they were not trained to account for hemodynamic stability. While the third agent includes MAP in its objective, this is reasonable in this case as well, since the agent was not exposed to significant fluctuations during training. This limitation probably arises because, in the maintenance phase, patients typically have a MAP around 70 mmHg or lower, and any additional drug infusion further reduces MAP, preventing the agent from exploring higher values (e.g., 30 units above baseline).
The policy employed by the third agent (Table 13), which is consistent with the one used in regular conditions, likely explains why it performs the worst when handling perturbations in the depth of hypnosis during disturbances. Specifically, it tends to lower the WAV_{cns} beyond the threshold compared to the other controllers while making minimal adjustments to MAP. The fact that the PID and Agent 2 strategies perform better while using the most remifentanil suggests that, although Agent 3's drug prioritization technique works well in controlled conditions, it may require modifications to manage disturbances effectively.

8.4. Explaining RL Agents Behavior

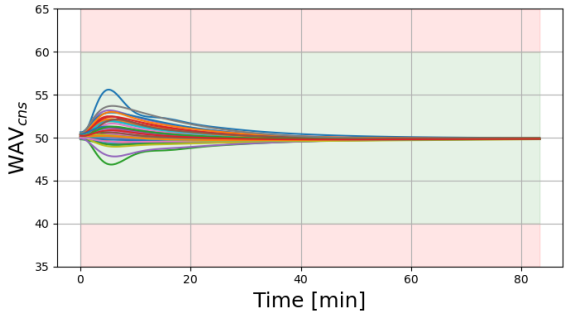
This section provides insight into the explainability of the RL agent's actions, focusing on Agent 2 and Agent 3, whose behavior is conditioned on demographic data. Understanding the decision-making process of RL agents is critical for ensuring reliability, transparency, and trust. The simulator used to train the agents is based on PK and PD models, specifically the Eleveld PK-PD models, which incorporate patient demographics such as age, weight, height, and gender to compute physiological responses to drug administration. One of the key characteristics of these models is that they introduce structured relationships between demographic factors and drug response, creating consistent patterns that guide the agent's learning process. However, while this structure allows us to evaluate whether the agent's learned policy aligns with the underlying PK/PD model, it also represents a limitation, as it has been shown that real-world physiological responses can deviate from PK-PD models predictions.
We employ SHapley Additive exPlanations (SHAP) [47] to interpret the agent's learned behavior. A SHAP value quantifies the contribution of each input feature to an action prediction. The SHAP value for feature i, denoted as ϕ_i, represents the change in the output when including that feature. The total model prediction can be written as the sum of the baseline prediction and the contributions of each individual feature:

$$f_x = \phi_0 + \sum_{i=0}^{n} \phi_i,$$

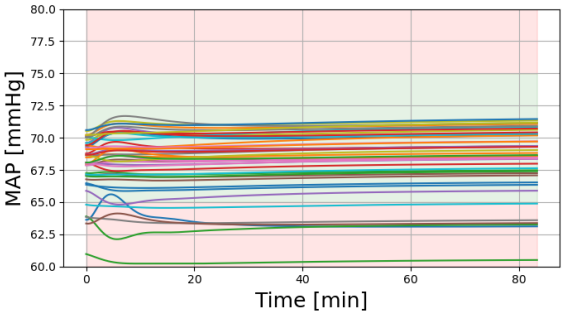
where f_x is the model's prediction for a specific input, and ϕ_0 represents the model's expected output when no features are provided.
Mathematically, the SHAP value ϕ_i is computed as the expected marginal contribution of feature i averaged over all possible feature combinations:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(f_x(S \cup \{i\}) - f_x(S) \right), \tag{24}$$

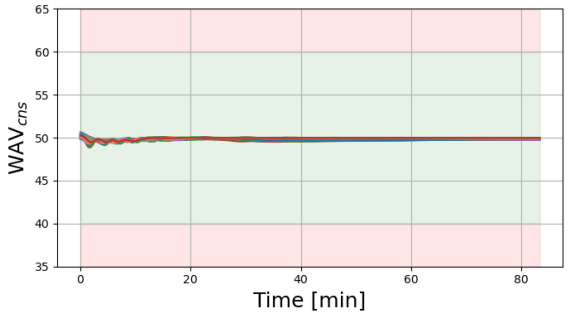
where N is the set containing all features, S is any subset of features not containing i, and $f_x(S)$ is the model's prediction when only the features in S are considered.
When computing the marginal contribution of feature i, given by $f_x(S \cup \{i\}) - f_x(S)$, if a feature is not included in S (or in $S \cup \{i\}$ when evaluating the contribution of feature i), its value is either omitted or replaced with a value drawn from a background dataset that represents the distribution of input states.
We compute SHAP values using the ExactExplainer from the SHAP library. This explainer is model-agnostic, meaning it operates independently of the model architecture.
Despite the computational complexity of this approach, we can afford to compute exact values as we only have 5 features for Agent 2 and 6 features for Agent 3. As clarified earlier, the ExactExplainer requires a background dataset that we build by sampling state observations from 100 simulations of in-silico patients. Each subset S in Equation (24) represents a different combination of features that are kept in their original state. For each test sample (i.e., one observation state from each of the 100 simulations), the ExactExplainer generates perturbed samples by considering every possible combination of features being replaced with values drawn from the background dataset. This means that if the patient state consists of features $[\text{Age}, \text{Height}, \text{Weight}, \text{Gender}, \text{WAV}_{\text{error}}, \text{MAP}_{\text{error}}]$, one such subset S might include only $[\text{Age}, \text{Weight}]$. In the corresponding perturbed sample, the values for the features not in S (i.e., $\text{Height}, \text{Gender}, \text{WAV}_{\text{error}}, \text{MAP}_{\text{error}}$) are replaced with typical background values.



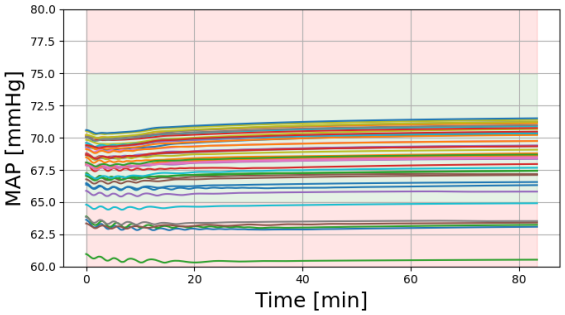
(a) WAV$_{\text{cns}}$ - PID



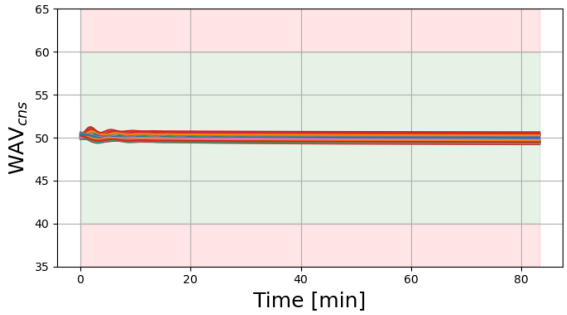
(b) MAP - PID



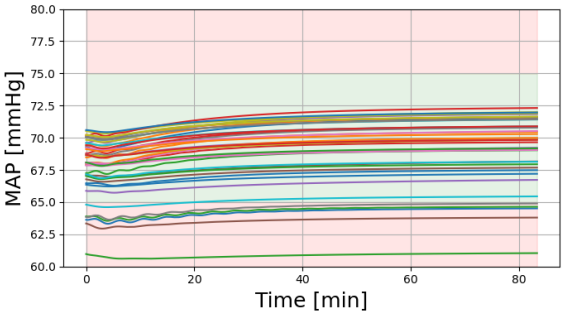
(c) WAV$_{\text{cns}}$ - Agent 1



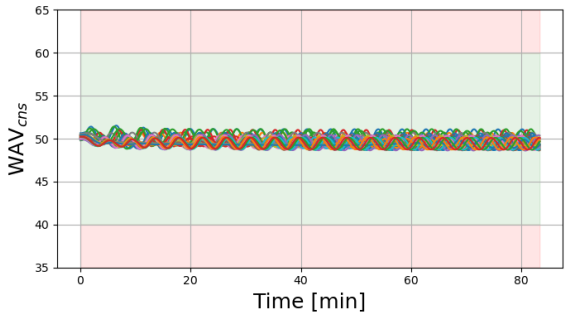
(d) MAP - Agent 1



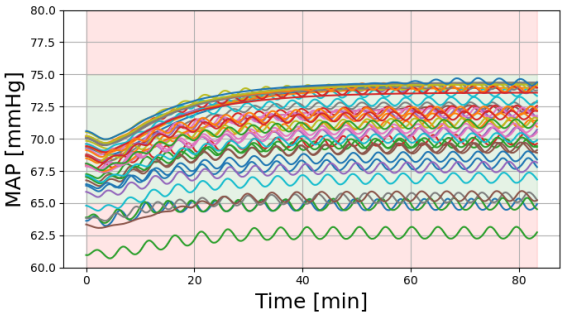
(e) WAV$_{\text{cns}}$ - Agent 2



(f) MAP - Agent 2



(g) WAV$_{\text{cns}}$ - Agent 3

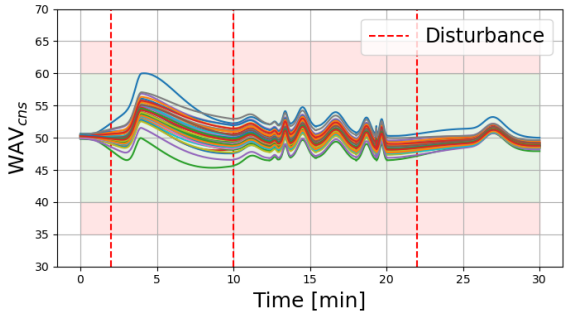


(h) MAP - Agent 3

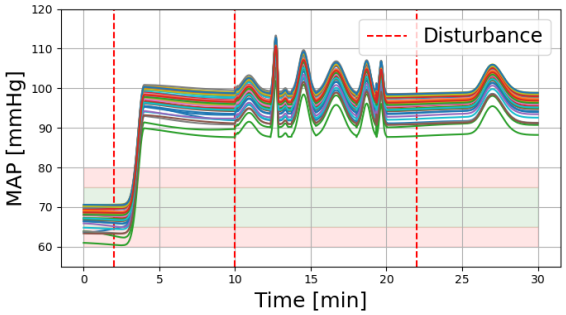
Figure 14: WAV and MAP results of control policies applied to 44 patients.

Each perturbed sample is then given the action prediction function to calculate the difference $f_x(S \cup \{i\}) - f_x(S)$. Finally, these marginal contributions are weighted and averaged over all possible subsets S following Equation (24).

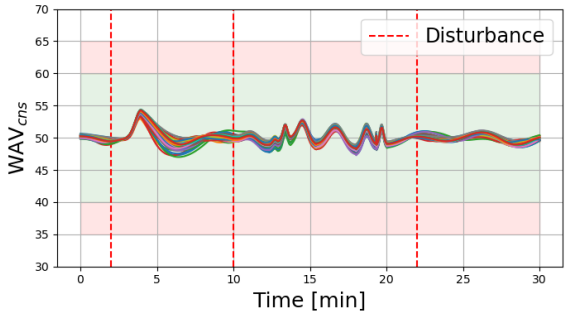
Results Figures 18 and 19 show the SHAP values for agents 2 and 3, respectively. Specifically, the subplots on the left (Figure 18a and Figure 19a) illustrate how each feature contributes to the decision regarding propofol infusion rates, while the subplots on the right display the feature importance for the remifentanil action. In



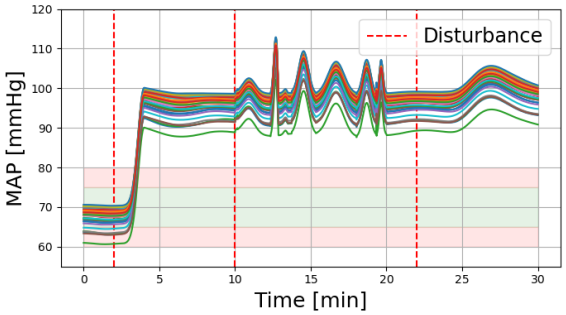
(a) WAV$_{\text{cns}}$ - PID



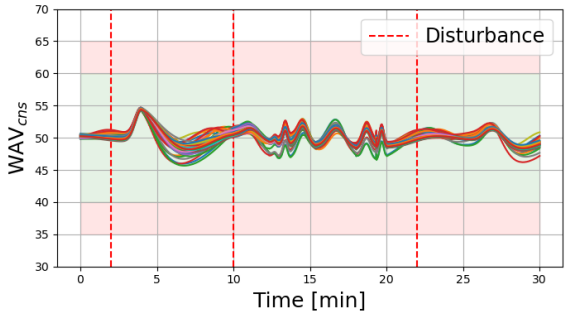
(b) MAP - PID



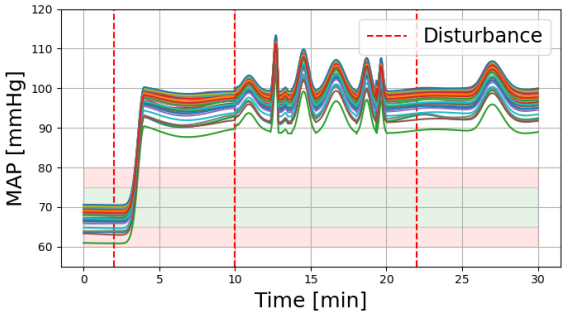
(c) WAV$_{\text{cns}}$ - Agent 1



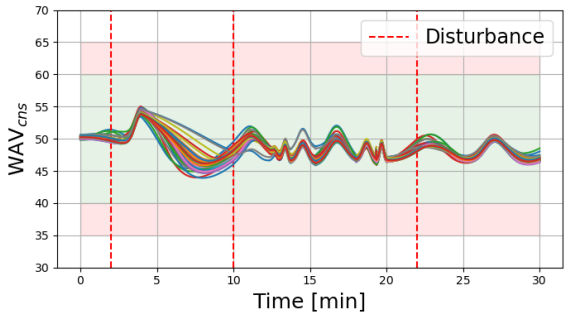
(d) MAP - Agent 1



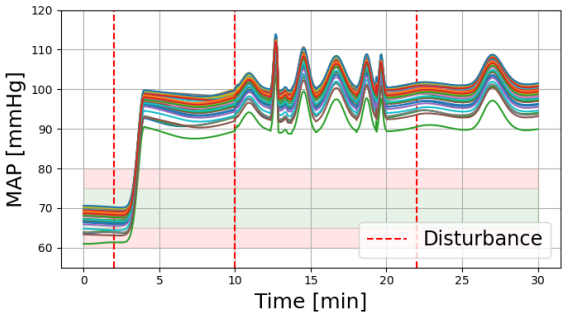
(e) WAV$_{\text{cns}}$ - Agent



(f) MAP - Agent 2



(g) WAV$_{\text{cns}}$ - Agent 3

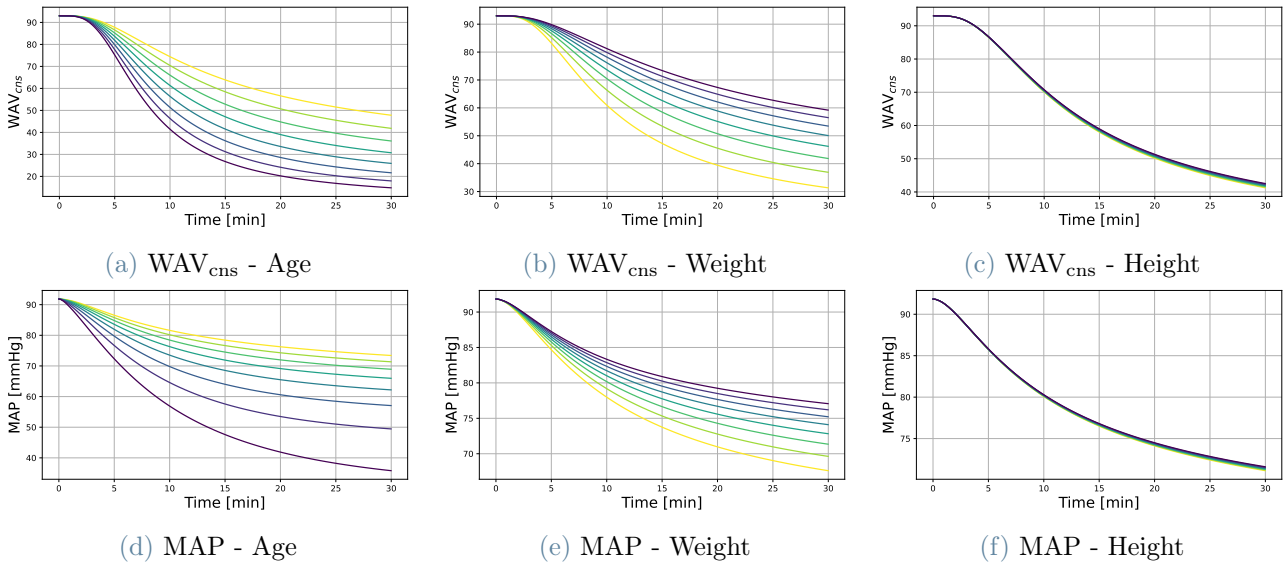


(h) MAP - Agent 3

Figure 15: WAV and MAP results of control policies applied to 44 patients, with added disturbances.

.

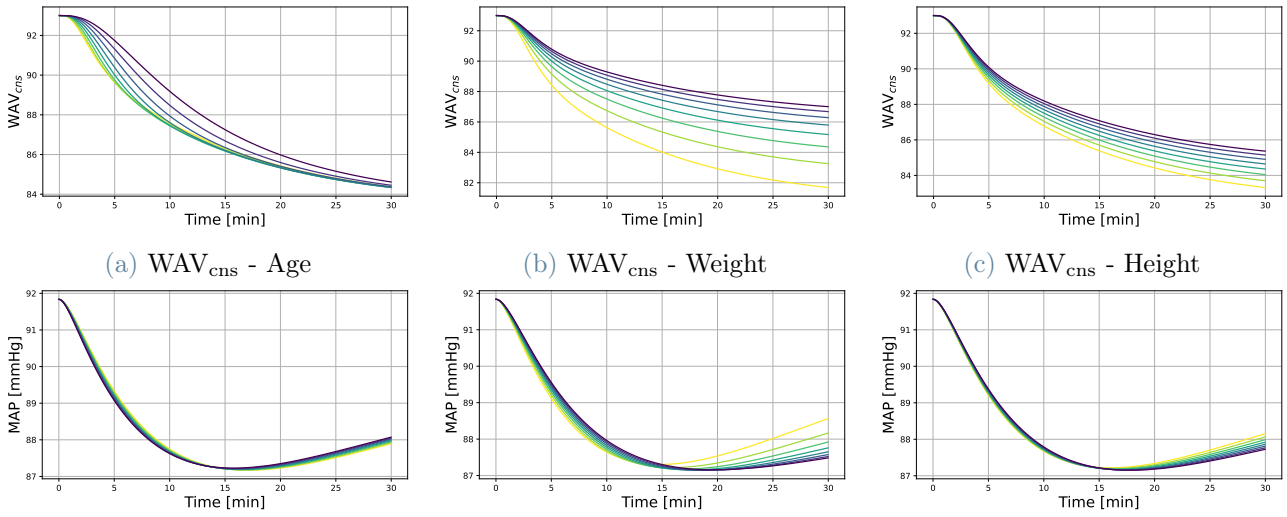
these plots, red dots indicate states where a feature has high values, whereas blue dots denote low values. The shift of dots to the right or left of the 0 value on the x-axis indicates the corresponding increase or decrease in the action due to changes in the input feature, relative to the expected model prediction when no features are considered. We remark that the order of features on the y-axis in the summary plot represents their importance in the model's predictions. Figures 16 and 17 illustrate the influence of demographic variables on WAV$_{\text{cns}}$ and MAP to help interpret the decision-making. Figure 16 shows how these parameters vary with age, weight, and height under a fixed infusion rate of 0.2 mg/s of propofol, while Figure 17 follows the same approach using 0.2



(a) WAV_{cns} - Age (b) WAV_{cns} - Weight (c) WAV_{cns} - Height

(d) MAP - Age (e) MAP - Weight (f) MAP - Height

Figure 16: Impact of Patient Demographic on WAV_{cns} and MAP During Propofol Administration.



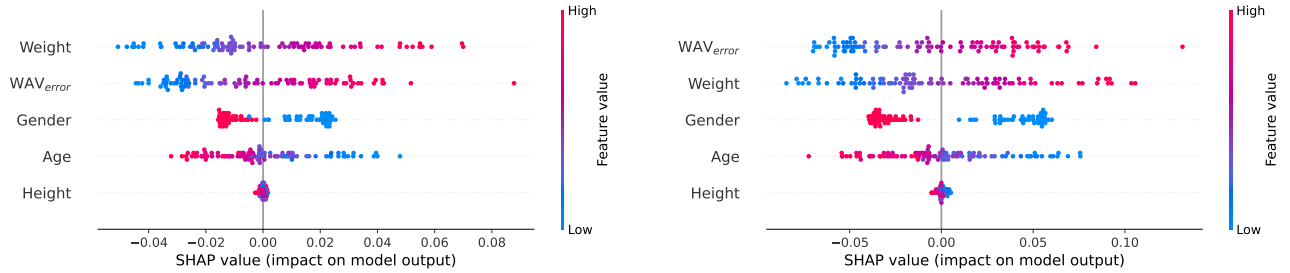
(a) WAV_{cns} - Age (b) WAV_{cns} - Weight (c) WAV_{cns} - Height

(d) MAP - Age (e) MAP - Weight (f) MAP - Height

Figure 17: Impact of Patient Demographic on WAV_{cns} and MAP During Remifentanil Administration.

µg/s of remifentanil. Although gender influences both WAV_{cns} and MAP values, the corresponding figures were omitted because the differences between male and female patients are not highly pronounced. However, we observe that when considering only propofol, female patients generally require slightly higher dosages than male patients to achieve similar therapeutic effects. For remifentanil, the behavior is the opposite. In both figures, warmer colors (e.g., yellow) indicate lower values for the corresponding variable. For example, in Figure 16b, WAV_{cns} values decrease for individuals with lower body weight.

Discussion We analyze the decision-making behavior of the agents, starting with Agent 2. For propofol administration, the features influencing the process, in order, are weight, WAV_{error}, gender, age, and height. The agent increases the action value with higher weight and WAV_{error}, while it decreases the action for female patients and older individuals (see Figure 18a). Height has negligible influence. Similarly, for remifentanil administration, the features influence the process in the order of WAV_{error}, weight, gender, age, and height, with an almost identical distribution. While it is straightforward to understand why a positive WAV_{error} increases the actions, we observe that the other variables also influence decision-making almost as expected. In particular, the agent's use of age and weight features aligns with real-world expectations. Older patients and individuals with higher weight (represented by red dots) receive reduced and increased propofol and remifentanil administration, respectively, which is consistent with clinical practice and the Eleveld models (see Figure 16a and Figure 17a for age, and Figure 16b and Figure 17b for weight).
For Agent 3 (Figure 19), the decision-making order differs. Figure 19a shows that the most influential feature for



(a) Propofol (b) Remifentanil

Figure 18: SHAP Values for Propofol and Remifentanil Actions – Agent 2.

(a) Propofol (b) Remifentanil

Figure 19: SHAP Values for Propofol and Remifentanil Actions – Agent 3.

propofol administration is $\text{WAV}_{\text{error}}$ followed by age, weight, height, $\text{MAP}_{\text{error}}$, and gender. The agent increases the action for higher $\text{WAV}_{\text{error}}$, weight, and height, while it decreases the action for older individuals. Notably, this time, gender has minimal influence. Since the agent frequently encounters states where $\text{MAP}_{\text{error}}$ is zero due to the dead zone defined in Section 4, its impact is almost negligible. This can be observed in the clustered red dots near the baseline value. However, when $\text{MAP}_{\text{error}}$ decreases, the agent reduces propofol administration accordingly.

Figure 19b shows the feature importance order for remifentanil, which is $\text{WAV}_{\text{error}}$, $\text{MAP}_{\text{error}}$, age, weight, height, and gender. The trends essentially mirror those of propofol. $\text{MAP}_{\text{error}}$ follows the same pattern as in propofol administration. Interestingly, while $\text{MAP}_{\text{error}}$ has almost no effect on propofol administration, it is the second most influential factor in the remifentanil action. This may be explained by the fact that propofol has a much stronger effect on WAV than on MAP, whereas remifentanil affects both parameters more equally.

While the relationships with $\text{WAV}_{\text{error}}$, age, weight, and $\text{MAP}_{\text{error}}$ are now clear (with $\text{WAV}_{\text{error}}$ and $\text{MAP}_{\text{error}}$ showing an obvious impact, and weight and age aligning with the PK/PD model), interpreting the influence of height is less straightforward. This complexity arises from height's relationship with weight through the BMI, making it difficult to isolate the direct effect of height alone.

We note that, when examining Figure 16 and Figure 17 used to interpret the decision-making of an agent, we are disregarding the profound synergy between propofol and remifentanil, as well as the potential relationships that might emerge when combining different demographic data. Therefore, while these trends help us understand the general pattern, they may not be completely accurate.

Overall, these findings suggest that the RL agents successfully learn the relationships between patient demographics, drug infusion rates, and physiological effects on the DoH and hemodynamic stability, making RL a promising approach to enhance patient care by learning patient-specific treatments.

9. Discussion and Conclusions

This study explores how RL can be used to automate TIVA during the maintenance phase without accounting for changes in the patient's condition due to external stimuli or complications. The focus is on understanding whether RL-based agents can precisely regulate the depth of hypnosis and hemodynamic stability in a patient-specific manner. We develop three models with the objective of increasing complexity. The first agent optimizes the policy of a PID controller, adjusting propofol administration while using a fixed propofol-to-remifentanil ratio. The second agent removes this constraint and incorporates patient demographic data into the state-space representation. The third agent extends this approach by also controlling MAP. The results show that RL

agents effectively learn to control anesthesia. The proposed agents were trained and evaluated in a simulated environment. The first agent outperforms the PID controller by achieving a similar control strategy with better performance, reducing error while using the exact amount of drugs. The second and third agents, despite their more complex objectives, achieve comparable performance. Moreover, using explainability techniques, we demonstrate that their decision-making aligns with the PK-PD model they are trained on. The agents also exhibit robustness to surgical stimuli, even though they are not explicitly trained to handle them. However, while they can compensate for perturbations in WAV_{cns}, they struggle with MAP. This is expected, as the first two agents do not consider hemodynamic stability, and even the third agent, which includes it in its objective, is not exposed to extreme MAP variations during training. These results suggest that RL systems have the potential to improve anesthesia control by providing a more personalized and adaptive approach. If trained successfully, agents based on RL can optimize drug delivery by improving accuracy and reducing overall drug consumption. Automation can reduce the workload of physicians by addressing problems such as fatigue and burnout. However, significant challenges remain.

9.1. Limitations

Despite these results, some limitations remain. The main concern arises from the simulator used to train and evaluate the agents. Although significant advances have been made in patient modeling, particularly in Pharmacokinetics and Pharmacodynamics, these models often provide a better estimate of population behavior at the expense of individual characteristics. For example, while elderly patients often require lower doses, this is not universally true, and the model's dependence on such assumptions may limit its accuracy. Despite integrating 44 patient-specific models, these might not be sufficient to train an agent capable of generalizing over a wide range of characteristics. Furthermore, we compared the results predicted by our simulation with real-world clinical data in Section 8.1. However, the comparison dataset is small and does not include recordings of norepinephrine administration or PTC.

Although the following aspects do not directly impact the training of the models developed here, we highlight some limitations that must be considered when using AReS to develop and evaluate any anesthesia controller. During the development of the simulator, we found no research in the literature that mathematically quantifies the effect of norepinephrine plasma concentration on SV and CO. Moreover, we could not identify studies that model complete interactions between all drugs considered. In particular, the interaction between rocuronium and other drugs, as well as the surface interaction between norepinephrine, propofol, and remifentanil on hemodynamic variables, remain unaddressed. The proposed disturbance model assumes that the examined patient is young and healthy, as medications such as beta-blockers [87] and antihypertensives, including ACE inhibitors [49], can limit compensatory increases in HR and MAP by inhibiting the response to endogenous adrenergic stimulation. While some papers describe the behavior of BIS, MAP, and other parameters in response to surgical stimuli, we leave the choice of the maximum amplitude of the patient's response to the user. This decision is made because, as far as we know, there are no data or studies on patient responses to surgical stimuli that are not heavily influenced by the amount of anesthetic agents administered. AReS incorporates the effect of non-traumatic fluid loss (defined as less than 30% of total fluid volume) on MAP but does not currently model traumatic fluid loss, as the implemented PK-PD models are not applicable in such cases. Although disturbances are mainly described as additive terms in the literature, our approach, which incorporates both additive noise and multiplicative terms, might not fully capture their physiological impact. A more realistic approach would involve adjusting PK and PD parameters.

Beyond simulator constraints, the RL agents themselves have limitations. Because agents are evaluated in the same simulated environment used for training, their real-world effectiveness remains unknown. Furthermore, the use of a deterministic policy presents challenges, particularly in exploration during training. The third agent, which aims to regulate both hypnosis and hemodynamic stability, struggles when MAP exceeds the safe range of 65–75 mmHg. Furthermore, these agents lack any built-in safety mechanisms to prevent harmful drug administrations, making their direct clinical application risky without further validation.

When considering the overall objective, these RL agents still lack key aspects of anesthesia control. They focus solely on maintaining anesthesia while induction is equally important and closely related to several postoperative risks. Furthermore, our approach does not account for the control of vasoconstriction, which is widely used to maintain a safe MAP, nor does it address the administration of neuromuscular blockade (NMB).

9.2. Future Developments

Future works should focus on refining the PK-PD models, particularly by improving the representation of interactions between drugs and different physiological variables. In addition, the impact of surgical disturbances should be characterized by modifying PK and PD parameters to create more realistic simulations. The collection

of large data sets containing complete clinical data would help to evaluate the accuracy of simulators such as AReS while also providing more realistic training and evaluation environments for RL agents. In addition, simulations during the training process should include disturbances, allowing them to adapt to a wider range of scenarios, thus improving robustness. Incorporating explicit signals indicating the onset of disturbances into the state representation could promote learning by enabling the models to recognize and reject them. Since disturbance-free and perturbed conditions may require different strategies, an alternative solution might involve training two specialized agents, one optimized for normal conditions and another designed one to handle disturbances. Collaboration with clinicians can be valuable in improving the training process, particularly in defining what constitutes safe or unsafe decisions, leading to more clinically relevant reward functions. In addition, the use of constrained RL approaches could help ensure that the learned policies remain within safe boundaries. This, coupled with the application of RL as a decision support system, in which agents provide recommendations rather than making autonomous decisions, could allow anesthesiologists to validate the model's suggestions without safety risks.

References

[1] Absalom, A. R. and Mason, K. P. (2017). *Total intravenous anesthesia and target controlled infusions*. Springer.

[2] Al-Rifai, Z. and Mulvey, D. (2016). Principles of total intravenous anaesthesia: basic pharmacokinetics and model descriptions. *BJA Education*, 16(3):92–97.

[3] Ali, H. H., Utting, J., and Gray, T. C. (1971). Quantitative assessment of residual antidepolarizing block (part I). *British Journal of Anaesthesia*, 43(5):473–477.

[4] Andreis, D. T. and Singer, M. (2016). Catecholamines for inflammatory shock: a Jekyll-and-Hyde conundrum. *Intensive Care Medicine*, 42(9):1387–1397.

[5] Andrzejowski, J., Sleigh, J. W., Johnson, I. A. T., and Sikiotis, L. (2000). The effect of intravenous epinephrine on the bispectral index and sedation. *Anaesthesia*, 55(8):761–763.

[6] Armstrong, S. A. and Herr, M. J. (2024). Physiology, Nociception. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

[7] Åström, K. and Hägglund, T. (1995). *PID Controllers: Theory, Design, and Tuning*. ISA - The Instrumentation, Systems and Automation Society.

[8] Aubouin-Pairault, B., Fiacchini, M., and Dang, T. (2023). Pas: a Python anesthesia simulator for drug control. *Journal of Open Source Software*, 8(88):5480.

[9] Bailey, J. M. and Shafer, S. L. (1991). A simple analytical solution to the three-compartment pharmacokinetic model suitable for computer-controlled infusion pumps. *IEEE Transactions on biomedical engineering*, 38(6):522–525.

[10] Bibian, S., Dumont, G. A., and Zikov, T. (2011). Dynamic behavior of bis, m-entropy and neurosense brain function monitors. *Journal of clinical monitoring and computing*, 25:81–87.

[11] Bibian, S. and Zikov, T. (2018). Neurosense® monitor with wavcns cortical quantifier: A deterministic approach to eeg analysis.

[12] Bibian, S., Zikov, T., Dumont, G., Ries, C., Puil, E., Ahmadi, H., Huzmezan, M., and MacLeod, B. (2001). Estimation of the anesthetic depth using wavelet analysis of electroencephalogram. In *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 951–955, Istanbul, Turkey. IEEE.

[13] Bouillon, T. W., Bruhn, J., Radulescu, L., Andresen, C., Shafer, T. J., Cohane, C., and Shafer, S. L. (2004). Pharmacodynamic interaction between propofol and remifentanil regarding hypnosis, tolerance of laryngoscopy, bispectral index, and electroencephalographic approximate entropy. *The Journal of the American Society of Anesthesiologists*, 100(6):1353–1372.

[14] Brady, M. C., Kinn, S., Stuart, P., and Ness, V. (2003). Preoperative fasting for adults to prevent perioperative complications. *Cochrane Database of Systematic Reviews*.

[15] Braithwaite, H. E., Payne, T., Duce, N., Lim, J., McCulloch, T., Loadsman, J., Leslie, K., Webster, A. C., Gaskell, A., and Sanders, R. D. (2023). Impact of female sex on anaesthetic awareness, depth, and emergence: a systematic review and meta-analysis. *British Journal of Anaesthesia*, 131(3):510–522.

[16] Brogi, E., Cyr, S., Kazan, R., Giunta, F., and Hemmerling, T. M. (2017). Clinical Performance and Safety of Closed-Loop Systems: A Systematic Review and Meta-analysis of Randomized Controlled Trials. *Anesthesia & Analgesia*, 124(2):446–455.

[17] Brown, E. N., Lydic, R., and Schiff, N. D. (2010). General Anesthesia, Sleep, and Coma. *New England Journal of Medicine*, 363(27):2638–2650.

[18] Calvi, G., Manzoni, E., and Rampazzo, M. (2022). Reinforcement Q-learning for Closed-loop Hypnosis Depth Control in Anesthesia. In *2022 30th Mediterranean Conference on Control and Automation (MED)*, pages 164–169, Vouliagmeni, Greece. IEEE.

[19] Coleman, R. M., Tousignant-Laflamme, Y., Gélinas, C., Choinière, M., Atallah, M., Parenteau-Goudreault, E., and Bourgault, P. (2013). Changes in the Bispectral Index in Response to Experimental Noxious Stimuli in Adults under General Anesthesia. *ISRN Pain*, 2013:1–8.

[20] Cooper, J., Newbower, R., Long, C., and McPeek, B. (1978). Preventable Anesthesia Mishaps: A Study of Human Factors. *Anesthesiology*, 49(6):399–406.

[21] Couto, M., Vide, S., Marco-Ariño, N., Troconiz, I. F., Esteves, S., Nunes, C. S., Amorim, P., and Mendes, J. (2022). Comparison of two pharmacokinetic–pharmacodynamic models of rocuronium bromide during profound neuromuscular block: analysis of estimated and measured post-tetanic count effect. *British Journal of Anaesthesia*, 128(3):473–481.

[22] Critchley, L. A. and Critchley, J. A. (1999). A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques. *Journal of clinical monitoring and computing*, 15:85–91.

[23] De Haes, A., Proost, J. H., Kuks, J. B., van den Tol, D. C., and Wierda, J. M. K. (2002). Pharmacokinetic/pharmacodynamic modeling of rocuronium in myasthenic patients is improved by taking into account the number of unbound acetylcholine receptors. *Anesthesia & Analgesia*, 95(3):588–596.

[24] Eleveld, D., Colin, P., Absalom, A., and Struys, M. (2018). Pharmacokinetic–pharmacodynamic model for propofol for broad application in anaesthesia and sedation. *British journal of anaesthesia*, 120(5):942–959.

[25] Eleveld, D. J., Proost, J. H., Vereecke, H., Absalom, A. R., Olofsen, E., Vuyk, J., and Struys, M. M. (2017). An allometric model of remifentanil pharmacokinetics and pharmacodynamics. *Anesthesiology*, 126(6):1005–1018.

[26] Fellahi, J.-L., Futier, E., Vaisse, C., Collange, O., Huet, O., Loriau, J., Gayat, E., Tavernier, B., Biais, M., Asehnoune, K., Cholley, B., and Longrois, D. (2021). Perioperative hemodynamic optimization: from guidelines to implementation—an experts' opinion paper. *Annals of Intensive Care*, 11(1):58.

[27] Fritz, B. A., Kalarickal, P. L., Maybrier, H. R., Muench, M. R., Dearth, D., Chen, Y., Escallier, K. E., Ben Abdallah, A., Lin, N., and Avidan, M. S. (2016). Intraoperative Electroencephalogram Suppression Predicts Postoperative Delirium. *Anesthesia & Analgesia*, 122(1):234–242.

[28] Fujimoto, S., Hoof, H. v., and Meger, D. (2018). Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35 th International Conference on Machine Learning*, volume 80, pages 1587–1596, Stockholm, Sweden.

[29] Gan, T., Soppitt, A., Maroof, M., El-Moalem, H., Robertson, K., Moretti, E., Dwane, P., and Glass, P. A. (2002). Goal-directed Intraoperative Fluid Administration Reduces Length of Hospital Stay after Major Surgery. *Anesthesiology*, 97(4):820–826.

[30] Gao, M., Liu, W., Chen, Z., Wei, W., Bao, Y., and Cai, Q. (2022). Global trends in anesthetic research over the past decade: a bibliometric analysis. *Annals of Translational Medicine*, 10(10):607.

[31] Godoy, A., Contreras, A., and Tabares, A. (2023). Agreement analysis of stroke volume and cardiac output measurement between a oscillometric device and transthoracic echocardiogram in normotensive individuals: a preliminary report. *Brazilian Journal of Anesthesiology*, 73(4):380–384.

[32] Gross, T., Feliot, E., Gayat, E., Chazot, T., Liu, N., Fischler, M., Bobet, M., and Le Guen, M. (2020). Bispectral Index During Maintenance of Total Intravenous Anesthesia: Frequency of Out of Recommended Range and Impact of Patients' Characteristics: A Brief Report. *Anesthesia & Analgesia*, 131(2):e52–e54.

[33] Hebbes, C. and Thompson, J. (2018). Pharmacokinetics of anaesthetic drugs at extremes of body weight. *BJA Education*, 18(12):364–370.

[34] Hemmerling, T. M. (2020). Robots will perform anesthesia in the near future. *Anesthesiology*, 132(2):219–220.

[35] Heusden, K. V., Ansermino, J. M., and Dumont, G. A. (2018). Robust miso control of propofol-remifentanil anesthesia guided by the neurosense monitor. *IEEE Transactions on Control Systems Technology*, 26:1758–1770.

[36] Hirose, M., Kobayashi, Y., Nakamoto, S., Ueki, R., Kariya, N., and Tatara, T. (2018). Development of a hemodynamic model using routine monitoring parameters for nociceptive responses evaluation during surgery under general anesthesia. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 24:3324.

[37] Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. Complex adaptive systems. MIT Press, Cambridge, Mass, 1st mit press ed edition.

[38] Hosseinirad, S., van Heusden, K., and Dumont, G. A. (2024). Evaluating inter-individual variability captured by the Eleveld pharmacokinetics model. *Journal of Clinical Monitoring and Computing*, 38(2):505–518.

[39] Ionescu, C. M., Neckebroek, M., Ghita, M., and Copot, D. (2021). An open source patient simulator for design and evaluation of computer based multiple drug dosing control for anesthetic and hemodynamic variables. *IEEE Access*, 9:8680–8694.

[40] Joachim, J., Cartailler, J., Vallee, F., Lefevre, T., Callebert, J., Gayat, E., and Lavielle, M. (2024). Design of a pharmacokinetic/pharmacodynamic model for administration of low dose peripheral norepinephrine during general anaesthesia. *British Journal of Clinical Pharmacology*, 90(11):2861–2869.

[41] Johnson, K. and Egan, T. D. (2015). Blood loss: does it change my intravenous anesthetic? In Smith, C. E., editor, *Trauma Anesthesia*, pages 105–112. Cambridge University Press, 2 edition.

[42] Johnston, A. J., Steiner, L. A., O'Connell, M., Chatfield, D. A., Gupta, A. K., and Menon, D. K. (2004). Pharmacokinetics and pharmacodynamics of dopamine and norepinephrine in critically ill head-injured patients. *Intensive care medicine*, 30:45–50.

[43] Kearse, L., Manberg, P., Chamoun, N., deBros, F., and Zaslavsky, A. (1994). Bispectral Analysis of the Electroencephalogram Correlates with Patient Movement to Skin Incision during Propofol/Nitrous Oxide Anesthesia. *Anesthesiology*, 81(6):1365–1370.

[44] Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[45] Lamke, L. O., Nilsson, G. E., and Reithner, H. L. (1979). Water loss by evaporation from the abdominal cavity during surgery. *Survey of Anesthesiology*, 23(4):222.

[46] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016). Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICRL 2016)*, Caribe Hilton, San Juan, Puerto Rico.

[47] Lundberg, S. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

[48] Mayne, D., Rawlings, J., Rao, C., and Scokaert, P. (2000). Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814.

[49] McConachie, I. and Healy, T. E. (1989). ACE inhibitors and anaesthesia. *Postgraduate Medical Journal*, 65(763):273–274.

[50] McGrath, C. D. and Hunter, J. M. (2006). Monitoring of neuromuscular block. *Continuing Education in Anaesthesia Critical Care & Pain*, 6(1):7–12.

[51] Meng, L., Yu, W., Wang, T., Zhang, L., Heerdt, P. M., and Gelb, A. W. (2018). Blood Pressure Targets in Perioperative Care: Provisional Considerations Based on a Comprehensive Literature Review. *Hypertension*, 72(4):806–817.

[52] Merigo, L., Padula, F., Latronico, N., Paltenghi, M., and Visioli, A. (2019). Optimized PID control of propofol and remifentanil coadministration for general anesthesia. *Communications in Nonlinear Science and Numerical Simulation*, 72:194–212.

[53] Mertens, M., Olofsen, E., Engbers, F. M., Burm, A. L., Bovill, J., and Vuyk, J. (2003). Propofol Reduces Perioperative Remifentanil Requirements in a Synergistic Manner. *Anesthesiology*, 99(2):347–359.

[54] Minto, C. F., Schnider, T. W., and Shafer, S. L. (1997). Pharmacokinetics and pharmacodynamics of remifentanil. ii. model application. *Anesthesiology*, 86(1):24–33.

[55] Monk, T. G., Saini, V., Weldon, B. C., and Sigl, J. C. (2005). Anesthetic Management and One-Year Mortality After Noncardiac Surgery:. *Anesthesia & Analgesia*, 100(1):4–10.

[56] Monnet, X., Jabot, J., Maizel, J., Richard, C., and Teboul, J. (2011). Norepinephrine increases cardiac preload and reduces preload dependency assessed by passive leg raising in septic shock patients. *Crit Care Med*, 4(39):689–694.

[57] Moore, B. L., Pyeatt, L. D., Kulkarni, V., Panousis, P., Padrez, K., and Doufas, A. G. (2014). Reinforcement Learning for Closed-Loop Propofol Anesthesia: A Study in Human Volunteers. *Journal of Machine Learning Research*, 15:655–696.

[58] Oron, A. P., Souter, M. J., and Flournoy, N. (2022). Understanding Research Methods: Up-and-down Designs for Dose-finding. *Anesthesiology*, 137(2):137–150.

[59] Osterman, J. E. and Van Der Kolk, B. A. (1998). Awareness during anesthesia and posttraumatic stress disorder. *General Hospital Psychiatry*, 20(5):274–281.

[60] Padmanabhan, R., Meskin, N., and Haddad, W. M. (2014). Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 1–8, Orlando, FL, USA. IEEE.

[61] Paravati, S., Rosani, A., and Warrington, S. J. (2024). Physiology, Catecholamines. In *StatPearls*. StatPearls Publishing.

[62] Pawlowski, A., Schiavo, M., Latronico, N., Paltenghi, M., and Visioli, A. (2023). Experimental Results of an MPC Strategy for Total Intravenous Anesthesia. *IEEE Access*, 11:32743–32751.

[63] Pawłowski, A., Schiavo, M., Latronico, N., Paltenghi, M., and Visioli, A. (2022). Model predictive control using MISO approach for drug co-administration in anesthesia. *Journal of Process Control*, 117:98–111.

[64] Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. (2018). FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[65] Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. Wiley-Interscience.

[66] Saldien, V., Vermeyen, K., and Wuyts, F. (2003). Target-controlled infusion of rocuronium in infants, children, and adults: a comparison of the pharmacokinetic and pharmacodynamic relationship. *Anesthesia & Analgesia*, 97(1):44–49.

[67] Sandre-Hernandez, O., Ramírez-Barrios, M., Ordaz, P., and Mera, M. (2023). Multivariable discrete MPC with exponential cost function on the automation of the anesthesia process. *International Journal of Robust and Nonlinear Control*.

[68] Schamberg, G., Badgeley, M., Meschede-Krasa, B., Kwon, O., and Brown, E. N. (2022). Continuous action deep reinforcement learning for propofol dosing during general anesthesia. *Artificial Intelligence in Medicine*, 123.

[69] Schnider, T. W., Minto, C. F., Gambus, P. L., Andresen, C., Goodale, D. B., Shafer, S. L., and Youngs, E. J. (1998). The influence of method of administration and covariates on the pharmacokinetics of propofol in adult volunteers. *The Journal of the American Society of Anesthesiologists*, 88(5):1170–1182.

[70] Schoffstall, J. M., Spivey, W. H., Davidheiser, S., Fuhs, L., and Kirkpatrick, R. J. (1990). Endogenous and exogenous plasma catecholamine levels in cardiac arrest in swine. *Resuscitation*, 19(3):241–251.

[71] Shafer, S. L. and Gregg, K. M. (1992). Algorithms to rapidly achieve and maintain stable drug concentrations at the site of drug effect with a computer-controlled infusion pump. *Journal of pharmacokinetics and biopharmaceutics*, 20(2):147–169.

[72] Short, T. G., Campbell, D., Frampton, C., Chan, M. T., Myles, P. S., Corcoran, T. B., Sessler, D. I., Mills, G. H., Cata, J. P., Painter, T., et al. (2019). Anaesthetic depth and complications after major surgery: an international, randomised controlled trial. *The Lancet*, 394(10212):1907–1914.

[73] Skytioti, M., Søvik, S., and Elstad, M. (2018). Respiratory pump maintains cardiac stroke volume during hypovolemia in young, healthy volunteers. *Journal of Applied Physiology*, 124(5):1319–1325.

[74] Smith, G., D'Cruz, J. R., Rondeau, B., and Goldman, J. (2025). General Anesthesia for Surgeons. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

[75] Snelder, N., Ploeger, B., Luttringer, O., Rigel, D., Fu, F., Beil, M., Stanski, D., and Danhof, M. (2014). Drug effects on the cvs in conscious rats: separating cardiac output into heart rate and stroke volume using pkpd modelling. *British Journal of Pharmacology*, 171(22):5076–5092.

[76] Soltesz, K., van Heusden, K., and Dumont, G. A. (2020). Models for control of intravenous anesthesia. In *Automated Drug Delivery in Anesthesia*, pages 119–166. Elsevier.

[77] Struys, M. M., De Smet, T., Glen, J. I. B., Vereecke, H. E., Absalom, A. R., and Schnider, T. W. (2016). The history of target-controlled infusion. *Anesthesia & Analgesia*, 122(1):56–69.

[78] Su, H., Koomen, J. V., Eleveld, D. J., Struys, M. M., and Colin, P. J. (2023). Pharmacodynamic mechanism-based interaction model for the haemodynamic effects of remifentanil and propofol in healthy volunteers. *British Journal of Anaesthesia*, 131(2):222–233.

[79] Sutton, R. S. and Barto, A. (2020). *Reinforcement learning: an introduction*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts London, England, second edition edition.

[80] Tsai, M. H. and Mangar, D. (2019). Decoupling Complexity in Perioperative Systems. *Anesthesia & Analgesia*, 129(5):e176–e176.

[81] Van Heusden, K., West, N., Umedaly, A., Ansermino, J. M., Merchant, R., and Dumont, G. (2014). Safety, constraints and anti-windup in closed-loop anesthesia. *IFAC Proceedings Volumes*, 47(3):6569–6574.

[82] Van Poucke, G. E., Bravo, L. J., and Shafer, S. L. (2004). Target controlled infusions: targeting the effect site while limiting peak plasma concentration. *IEEE transactions on biomedical engineering*, 51(11):1869–1875.

[83] Vandemoortele, O., Hannivoort, L. N., Vanhoorebeeck, F., Struys, M. M. R. F., and Vereecke, H. E. M. (2022). General Purpose Pharmacokinetic-Pharmacodynamic Models for Target-Controlled Infusion of Anaesthetic Drugs: A Narrative Review. *Journal of Clinical Medicine*, 11(9):2487.

[84] Varvel, J. R., Donoho, D. L., and Shafer, S. L. (1992). Measuring the predictive performance of computer-controlled infusion pumps. *Journal of pharmacokinetics and biopharmaceutics*, 20:63–94.

[85] Viby-Mogensen, J., Howardy-Hansen, P., Chræmmer-Jørgensen, B., Ørding, H., Engbæk, J., and Nielsen, A. (1981). Posttetanic Count (PTC): A New Method of Evaluating an Intense Nondepolarizing Neuromuscular Blockade. *Anesthesiology*, 55(4):458–461.

[86] Wanner, P. M., Wulff, D. U., Djurdjevic, M., Korte, W., Schnider, T. W., and Filipovic, M. (2021). Targeting Higher Intraoperative Blood Pressures Does Not Reduce Adverse Cardiovascular Events Following Noncardiac Surgery. *Journal of the American College of Cardiology*, 78(18):1753–1764.

[87] Warden, J. C. (1973). Beta-adrenergic blocking drugs in anaesthesia. *Anaesthesia and Intensive Care*, 1(4):289–292.

[88] Weiss, B. M., Spahn, D. R., Rahmig, H., Rohling, R., and Pasch, T. (1996). Radial artery tonometry: moderately accurate but unpredictable technique of continuous non-invasive arterial pressure measurement. *British Journal of Anaesthesia*, 76(3):405–411.

[89] West, N., van Heusden, K., Gorges, M., Brodie, S., Rollinson, A., Peterson, C., Dumont, G., Ansermino, J., and Merchant, R. (2018). Design and evaluation of a closed-loop anesthesia system with robust control and safety system. *Anesthesia and Analgesia*, 127:883–894.

[90] William Tank, A. and Lee Wong, D. (2014). Peripheral and central effects of circulating catecholamines. In Terjung, R., editor, *Comprehensive Physiology*, pages 1–15. Wiley, 1 edition.

[91] Yamada, K., Shingu, K., Tsushima, K., and Murao, K. (1997). Circulatory and catecholamine responses to tracheal intubation and skin incision during sevoflurane, isoflurane, or halothane anesthesia. *Journal of Anesthesia*.

[92] Yang, H., Deng, H.-M., Chen, H.-Y., Tang, S.-H., Deng, F., Lu, Y.-G., and Song, J.-C. (2022). The Impact of Age on Propofol Requirement for Inducing Loss of Consciousness in Elderly Surgical Patients. *Frontiers in Pharmacology*, 13:739552.

[93] Yun, W. J., Shin, M., Mohaisen, D., Lee, K., and Kim, J. (2024). Hierarchical Deep Reinforcement Learning-Based Propofol Infusion Assistant Framework in Anesthesia. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2510–2521.

[94] Yun, W. J., Shin, M. J., Jung, S., Ko, J. G., Lee, H. C., and Kim, J. (2023). Deep reinforcement learning-based propofol infusion control for anesthesia: A feasibility study with a 3000-subject dataset. *Computers in Biology and Medicine*, 156.

[95] Zhou, Y., Lindsey, B., Snyder, S., Bell, E., et al. (2024). Sampling rate requirement for accurate calculation of heart rate and its variability based on the electrocardiogram. *Physiological Measurement*, 45(2):025007.

[96] Zierle-Ghosh, A. and Jan, A. (2025). Physiology, Body Mass Index. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

Abstract in lingua italiana

Ogni anno, più di 300 milioni di interventi chirurgici in tutto il mondo richiedono l'anestesia generale. Tuttavia, nonostante la sua complessità, questo processo è ancora gestito manualmente. Per migliorare il benessere del paziente, i sistemi automatizzati ad anello chiuso sono stati esaminati e hanno dimostrato di migliorare la stabilità e la sicurezza del paziente. Questo studio presenta tre agenti di Reinforcement Learning (RL) per il controllo dell'anestesia durante la fase di mantenimento, addestrati e valutati con un simulatore di paziente da noi progettato per fornire un ambiente simulato. Il primo agente imita la politica di un controllore Proporzionale-Integrale-Derivativo (PID), utilizzato come riferimento, regolando l'anestetico per controllare la profondità dell'ipnosi in base a un rapporto fisso tra ipnotico e oppioide. Il secondo agente incorpora i dati demografici del paziente ed elimina il vincolo di un rapporto fisso. Il terzo agente estende l'obiettivo del secondo monitorando anche la stabilità emodinamica. Gli agenti di RL vengono confrontati con il controllore PID utilizzando metriche di controllo standard. Il primo agente dimostra prestazioni superiori nel controllo del sedativo e della profondità dell'ipnosi, mentre il secondo e il terzo agente hanno prestazioni comparabili a quelle del controllore PID, nonostante gestiscano compiti più complessi. Sebbene non siano esplicitamente addestrati per gestire i disturbi, tutti gli agenti sono in grado di compensare le perturbazioni nella profondità dell'ipnosi, ma hanno difficoltà con le deviazioni della pressione arteriosa media. L'analisi di explainability mostra che le politiche del secondo e del terzo agente sono in linea con i modelli farmacocinetici e farmacodinamici su cui sono stati allenati. Questi risultati suggeriscono che RL offre un approccio promettente per il controllo personalizzato dell'anestesia.

Parole chiave: controllo automatico dell'anestesia, reinforcement learning, modellizzazione farmacocinetica-farmacodinamica, software di simulazione del paziente, trattamento personalizzato del paziente