Journal of Machine Learning Research 15 (2014) 655-696 Submitted 9/12; Revised 12/13; Published 2/14

Reinforcement Learning for Closed-Loop Propofol Anesthesia: A Study in Human Volunteers

Brett L Moore BRETT.MOORE@IEEE.ORG
Department of Computer Science
Texas Tech University
Lubbock, TX 79409, USA

Larry D Pyeatt LARRY.PYEATT@SDSMT.EDU
Department of Mathematics and Computer Science
South Dakota School of Mines and Technology
Rapid City, SD 57701, USA

Vivekanand Kulkarni VKULKARNI@STANFORD.EDU
Periklis Panousis PANOUSIS@STANFORD.EDU
Kevin Padrez KPADREZ@GMAIL.COM
Anthony G Doufas AGDOUFAS@STANFORD.EDU
Department of Anesthesiology, Perioperative and Pain Medicine
Stanford University School of Medicine
Stanford, CA, 94305, USA

Editor: Peter Dayan

Abstract

Clinical research has demonstrated the efficacy of closed-loop control of anesthesia using the bispectral index of the electroencephalogram as the controlled variable. These controllers have evolved to yield patient-specific anesthesia, which is associated with improved patient outcomes. Despite progress, the problem of patient-specific anesthesia remains unsolved. A variety of factors confound good control, including variations in human physiology, imperfect measures of drug effect, and delayed, hysteretic response to drug delivery. Reinforcement learning (RL) appears to be uniquely equipped to overcome these challenges; however, the literature offers no precedent for RL in anesthesia. To begin exploring the role RL might play in improving anesthetic care, we investigated the method's application in the delivery of patient-specific, propofol-induced hypnosis in human volunteers. When compared to performance metrics reported in the anesthesia literature, RL demonstrated patient-specific control marked by improved accuracy and stability. Furthermore, these results suggest that RL may be considered a viable alternative for solving other difficult closed-loop control problems in medicine. More rigorous clinical study, beyond the confines of controlled human volunteer studies, is needed to substantiate these findings.

Keywords: reinforcement learning, bispectral index, propofol, anesthesia, hypnosis, closed-loop control

1. Introduction

When compared to standard population-based dosing, patient-specific drug administration is generally preferred in the clinical practice of anesthesia. Computer-controlled drug deliv-

ery systems have been investigated as a means of achieving patient-specific anesthesia(Liu et al., 2013, 2012; Hahn et al., 2011; Hemmerling et al., 2010), and their application is associated with a number of favorable patient outcomes, including decreased intraoperative drug consumption and shortened postoperative recovery times (Liu et al., 2006; Servin, 1998; Theil et al., 1993). Historically, the application of conventional control techniques, such as proportional-integral-derivative (PID) control, in closed-loop anesthesia has shown moderate success (Absalom and Kenny, 2003). However, these historical successes have been constrained by the PID method's inherent limitations, as well as the complexity of human physiology (Wood, 1989). To improve control performance, clinical study has broadened to include techniques commonly associated with intelligent systems, most notably Bayesian filtering and fuzzy control(Ching et al., 2013; Shanechi et al., 2013; De Smet et al., 2008; Esmaeili et al., 2008; Carregal et al., 2000; Schaublin et al., 1996).

Reinforcement learning (RL), one of many intelligent system techniques, has demonstrated proficiency in difficult robotic control tasks (Gullapalli, 1993). However, RL has no reported application to clinical control problems, with the exception of work leading to this study (Moore et al., 2011a,b, 2004). Nonetheless, RL has a presence in medicine, and reported applications include ultrasound image segmentation (Sahba et al., 2008) and planning tasks, such as scheduling of HIV therapy (Ernst et al., 2006), optimizing deep-brain stimulation in epilepsy treatment (Guez et al., 2008), dosing strategies for anemia management in patients with chronic renal failure (Martín-Guerrero et al., 2009; Gaweda et al., 2006), and clinical trial design (Zhao et al., 2009). These applications support the assertion that reinforcement learning can serve as a "medical decision aid" (Martín-Guerrero et al., 2009). However, RL's aptitude for specialized clinical application remains incompletely explored since these applications were non-clinical. In the examples cited, RL was applied to data collected from patients, but no RL algorithm contributed directly to patient care.

This lack of direct application does not imply that RL is unsuited for computer-controlled drug delivery since the method has been successfully applied to critical real-time industrial control tasks (Ernst et al., 2009). Furthermore, the basic principles of reinforcement learning (dynamic programming and value function optimization) have been studied in depth-of-anesthesia control with favorable results (Hu et al., 1994). Thus, the two-fold objectives of this study were to a) investigate the clinical suitability of reinforcement learning for closed-loop control of intravenous propofol anesthesia in healthy human volunteers, and b) compare the performance of RL control against published clinical metrics. To accomplish these objectives, an RL agent was developed, tested *in silico*, and then evaluated in healthy volunteers under an IRB-approved study protocol in the Stanford University School of Medicine Department of Anesthesiology, Perioperative, and Pain Medicine.

2. Background

To begin answering the question "why should reinforcement learning be applied in anesthesia," this section establishes the problem with an introduction to the motivation and challenges of closed-loop control of intraoperative hypnosis. Discussion continues by summarizing the manner in which RL can address deficiencies in some contemporary approaches.

2.1 Propofol-Induced Hypnosis

Propofol is a short-acting sedative agent administered intravenously to achieve induction and maintenance of general anesthesia in the operating room and other critical care arenas. Propofol suppresses higher brain function to produce *hypnosis*, a suppression of consciousness.[1] Propofol, like other hypnotic agents, achieves unconsciousness "by altering neurotransmission at multiple sites in the cerebral cortex, brain stem, and thalamus." (Brown et al., 2010, p. 2641). For a thorough treatment of propofol's mechanism of action, see Brown et al.

The anesthesia community has studied automated delivery of propofol-induced hypnosis, in part, because the drug and its pharmacodynamic effects satisfy basic requirements for closed-loop control. To accomplish such *feedback* control, a controller must first be equipped to a) influence the desired control parameter, and b) observe the affects of its actions on that control parameter. The short-acting nature of propofol, characterized by rapid onset and recovery, readily satisfies the first requirement (Vanlersberghe and Camu, 2008). The complexities of the human central nervous system and its interaction with propofol make objective, quantitative measurement of hypnosis (control effect) challenging, but—as the following sections show—measurement of propofol effect is feasible.

2.2 Depth of Hypnosis Measurement

This section introduces the use of the electroencephalogram and its derivatives in the assessment of hypnotic depth. Some of the challenges associated with these methods are also discussed.

2.2.1 Electroencephalogram (EEG)

In closed-loop regulation of hypnosis, the controlled variable is the patient's level of consciousness, or awareness. Cerebral electrical activity is correlated with consciousness (Brown et al., 2010), and hypnosis (suppression of awareness) displays as change in cortical electrical activity. Electroencephalography, the measurement of cerebral electrical activity, produces the *electroencephalogram* (EEG). The EEG is often obtained with an non-invasive array of scalp sensors. Normal, waking cortical electrical activity is marked by periodic signals in five narrow frequency bands, α, β, γ, δ, and θ. When an EEG is obtained transcutaneously, signals within these bands range in the tens of millivolts. Accurate capture of this low-power signal is complicated by non-cortical biologic artifacts: eye motion and blinking, facial muscle movement, and cardiac pulse (Fitzgibbon et al., 2007).[2] Other factors, like changes in skin conductance, can impact the fidelity of signal acquisition. The EEG is also susceptible to contamination from electrical sources found in the intraoperative environment: power lines, overhead lighting, electrocautery, and other medical devices. For these reasons, isolation and removal of non-cortical artifacts remains a challenging problem for EEG analysis and interpretation.

1. Hypnosis is just one member of a collection of clinical endpoints that comprise "general anesthesia"; others include akinesia (immobility), amnesia, analgesia, and autonomic system stability.
2. Electrooculography (EOG), electromyography (EMG) and electrocardiography (ECG) are the practices of measuring these "unwanted" signals.

Research has shown that when Fourier analysis (or another time-frequency analysis method) is applied to the EEG, the energy content within the five spectral bands can provide insight into depth of hypnosis since the spectral features of the EEG signal are modulated with level of consciousness. Studies demonstrate that the hypnotic component of general anesthesia "produces distinct patterns on the electroencephalogram (EEG), the most common of which is the progressive increase in low-frequency, high-amplitude activity as the level of general anesthesia deepens." (Brown et al., 2010, p. 2638).

2.2.2 Bispectral Analysis of the EEG

The practiced anesthetist can discern and interpret changes in EEG power spectra associated with hypnosis induction, maintenance, and emergence; however, the relationship of these spectral components (and their changes) to depth of hypnosis is not obvious. Thus, processed EEG variables have been studied with the aim of developing a "simplified interpretation of the EEG" for objective, broadly-applicable measures of anesthetic depth (Sigl and Chamoun, 1994, p.392). One such indicator, the bispectral index (Covidien, Mansfield, MA), is well-reported in the anesthesia literature. BIS, as it is known, differs from conventional quantitative EEG parameters in that it augments traditional power spectral (Fourier) methods with *bispectral analysis*, a means of measuring *phase coupling* between pairs of frequency components. This added dimension of *bicoherence* can improve identification of EEG patterns associated with varying levels of cortical activity.

2.2.3 The Bispectral Index of the EEG (BIS)

Sigl and Chamoun define the bispectral index of the EEG as "a multivariate measure incorporating bispectral and time-domain parameters derived from the EEG," (1994, p. 402). This proprietary index was developed by statistically linking the EEG's time- and frequency-domain features to a database of hand-selected "behavioral assessments of sedation and hypnosis," (Rampil, 1997, p. 998). The result, BIS, is a weighted sum of processed EEG features tied to the clinical endpoints of hypnosis that is "insensitive to the specific anesthetic or sedative agent," (Rampil, 1997, p. 1000). This bispectral index lies in the range [0, 100] (Sigl and Chamoun, 1994; Rampil, 1997). A measure of 100 is associated with normal wakefulness; a value of 0 correlates to an iso-electric brain state.[3]

Research has shown that evidence of propofol's pharmacodynamic effect may be observed in the bispectral index of the EEG: "The BIS both correlated well with the level of responsiveness and provided an excellent prediction of the loss of consciousness. These results imply that BIS may be a valuable monitor of the level of sedation and loss of consciousness for propofol, midazolam, and isoflurane." (Glass et al., 1997). This finding is consistent with expectation: BIS was developed to be a statistical correlation between EEG patterns and clinical attributes of hypnosis: loss of consciousness, progressive loss of reflexes, return of consciousness, etc.

However, BIS has been observed to be an imperfect indicator of hypnotic condition. Some of the challenges stem from noise contamination in the underlying EEG signal. For example, EMG signals, such as those resulting from eye or facial motion, may overlap the

3. Thorough treatments of bispectral analysis of the EEG may be found in Sigl and Chamoun (1994) and Rampil (1997).

EEG's higher frequency β and γ bands. This sort of EEG signal contamination has been associated with elevated BIS values in surgical patients (Renna et al., 2002). To attenuate the influence of electrical noise, the A-2000 BIS monitor, like the one used in this study, applies selective band-pass and low-pass digital filters in its process of computing BIS.

Research also indicates that contamination of the BIS signal extends beyond external electrical noise; normal physiologic processes can play a role. BIS and EMG variability, characterized by relatively high-frequency fluctuation, has been observed to predict somatic responses to noxious stimulus (Bloom et al., 2008; Greenwald and Rosow, 2006). In more recent work, measures of BIS and EMG variability were computed as standard deviations over a 3-min window of samples. The resulting sBIS and sEMG indicators predicted somatic response to painful stimuli (Mathews et al., 2012). The implication is meaningful to closed-loop control of hypnosis: BIS variability seems positively correlated to lack of analgesia, rather than hypnosis. Since propofol is not an analgesic agent, it's reasonable to conclude that high-frequency changes in BIS should not contribute to propofol delivery decisions. In acknowledgment of these issues, the A-2000 BIS monitor provides a user-selectable option that applies either a 15-sec or 30-sec smoothing window to its BIS measurements. The manual advises the user to select the smoothing windows according to a desire for "decreased delay" or "decreased variability."

Other research highlights additional sources of "noise" that may influence the BIS signal. Dahaba provides an excellent survey of clinical and physiological conditions that perturb BIS measurement (2005). In light of these factors, it's reasonable to consider BIS as a probabilistic indicator of hypnotic depth, not an absolute one. As such, probabilistic control methods, like RL, become increasingly relevant.

2.3 Motivation for Good Control of Hypnosis

BIS has been recently studied as a mitigation for risk of unintended intraoperative awareness, defined as conscious behavior (motion, vocalization, etc.) during surgery or post-operative recall of intraoperative events. Unintentional intraoperative awareness can challenge the anesthetist because doses ensuring adequate hypnosis may lead to hemodynamic and/or respiratory instabilities in sensitive patients (i.e., trauma, critically-ill, and elderly). While the incidence of intraoperative awareness is estimated to be low, 0.13% (Sebel et al., 2004), it can be severely traumatic for the patient. BIS monitoring has been recommended as a preventative measure (Sandin et al., 2000) and has been reported to reduce the incidence of unintended intraoperative awareness (Myles et al., 2000). This finding remains controversial since this evidence comes from observational clinical trials (Avidan et al., 2008), and the execution of a convincing prospective clinical trial is logistically difficult.

At first glance, the risk of unintended intraoperative awareness implies that "deeper is better." However, higher doses of propofol are correlated with respiratory and hemodynamic depression. Emerging research substantiates a balance in hypnosis with reports of a possible causal link between deep anesthesia (BIS < 45) and postoperative morbidity (Lindholm et al., 2009). Again, this conclusion requires further substantiation before wide-spread acceptance.

These opposing concerns, awareness versus toxicity, as well as the favorable outcomes cited previously, link good control of intraoperative anesthesia to good patient care. Conse-

quently, closed-loop control of propofol-induced hypnosis is well-represented in the literature (Liu et al., 2013, 2012; Hahn et al., 2011; Hemmerling et al., 2010; Struys et al., 2007, 2004; Absalom and Kenny, 2003; Leslie et al., 2002; Absalom et al., 2002; Sakai et al., 2000; Struys et al., 2001), yet accurate and stable control of intraoperative hypnosis remains an incompletely solved problem.

2.4 Challenges to Optimal Control of Hypnosis

Optimal control of propofol-induced hypnosis is a difficult problem for several reasons. Properties of the patient, the drug, and the intraoperative environment all contribute confounding influences. The patient's age, gender, and ethnicity, as well as disease and surgical intervention (Schnider et al., 1998; Barvais et al., 1996), are known to affect response to propofol infusion. Additionally, "intra-subject heterogeneity", or tendency for change in an individual (Rigby-Jones and Sneyd, 2012), assures that any accurate characterization of a patient's propofol response has a limited lifetime. For these reasons, commercially available target-controlled infusion (TCI) systems rely on general, population models of drug effect, leaving them unequipped for patient-specific drug delivery.

Additionally, a system regulating a patient's propofol concentration is limited to an *asymmetric* influence that further hinders good control. Propofol concentrations can be readily increased via intravenous infusion; however, the system lacks a direct means of decreasing concentration. Instead, the controller must wait for the patient to decrease propofol concentration through metabolism or redistribution. As a consequence, the controller possesses a direct means of increasing hypnosis, but an indirect means of decreasing hypnosis.

Other aspects of propofol infusion are problematic. The delay between action (infusion) and effect (hypnosis) can exceed two minutes. This delay (*transport delay* in control literature) is variable, hysteretic, and demonstrates flow rate dependence (Struys et al., 2007; Pilge et al., 2006). In addition, propofol's effect on consciousness is nonlinear, meaning that a fixed dose of propofol can impact BIS differently, depending on the patient's level of hypnosis at the time of infusion. As a result, the controller cannot always assume that a chosen dose will always have the same effect.

Finally, hypnosis is a balance of stimulus and drug effect. In the absence of stimulus, a relatively low concentration of propofol can yield the desired BIS. The onset of a routine surgical event (incision, manipulation, etc.) can disturb this equilibrium, rendering the patient's previously adequate concentration insufficient and leading to an undesired increase in consciousness. Thus, a clinically relevant hypnosis control system should be prepared to compensate for those external influences that can negatively impact control (Röpcke et al., 2001b; Ausems et al., 1986).

2.5 Conventional Control

Much of the initial progress in closed-loop anesthesia has been accomplished using conventional control techniques, like Proportional-Integral-Derivative (PID) control (Absalom et al., 2002; Struys et al., 2001; Sakai et al., 2000; Kenny and Mantzaridis, 1999; Mortier et al., 1998). These classical control methods enjoy widespread industrial application due to their simplicity of design and implementation, as well as their success in many control

problems. Furthermore, a measure of the PID technique's popularity is due to its foundation in classical control theory, its lack of dependence upon an accurate process model, and its ease in implementation.

Given the clinical interest in well-controlled hypnosis, it is no surprise that PID (along with its PI and PD variants) has been applied to hypnosis control. Kenny and Mantzaridis used a basic proportional-integral controller to regulate hypnosis in surgical patients (1999). This automated system delivered satisfactory anesthesia in a population of 100 patients and demonstrated that the hypnotic process, although noisy and uncertain, may be regulated using conventional techniques. Further research has demonstrated similar results: Absalom et al. coupled the bispectral index with a PID controller and observed largely satisfactory results in the administration of general anesthesia in ten patients (2002).

Despite instances of successful PID control in general anesthesia, the technique should not be applied universally with an expectation of similar results. Constant-coefficient PID methods, like those historically applied in hypnosis control, are not equipped to satisfactorily control processes with variable time delays, variable plant parameters, significant nonlinearities, and non-negligible process noise. Olkkola summarizes the use of PID in closed-loop control of anesthesia: "PID controllers are in general not universally applicable to nonlinear concentration-response curves..." (Olkkola et al., 1991, 420). Our simulation work supports this assertion (Moore, 2003). In general, a PID controller may be tuned to perform well for an arbitrary patient at a fixed level of hypnosis. However, the same controller would perform poorly when patient characteristics or hypnosis target varied. More convincingly, clinical observations support Olkkola's claim, as well. Absalom et al. observed oscillation around the BIS setpoint in the operating room (2002), and Leslie et al. observed similar oscillations in a conscious sedation experiment (2002).

Given the known limitations of the constant-coefficient PID controller, as well as the reported instances of sub-optimal control, it can be reasonably concluded that constant-coefficient PID is not the ideal solution (Struys et al., 2001). Current anesthesia literature suggests the ideal solution is a model-based, adaptive system(Ching et al., 2013; Shanechi et al., 2013). These systems do not exclude the PID class of controllers since neural networks, among other methods, have been used to establish relationships between system inputs and variable PID coefficients (Omatu et al., 1996). However, it should be noted that adding a model increases the complexity of the PID controller, thereby eroding its advantage of simplicity.

2.6 Reinforcement Learning

Reinforcement learning (RL) is an intelligent control method that provides a structured, mathematically robust mechanism for goal-directed decision making in which long-term gain is maximized (Sutton and Barto, 1998; Kaelbling et al., 1996). Unlike supervised learning methods, no examples of desired behavior are provided during training; instead, favorable action choices are encouraged through positive and/or negative reinforcements. Under this framework, knowledge is gained through experimentation: actions are chosen, effects are observed, and rewards are gained.

3. Methods

To begin assessing the suitability of reinforcement learning for closed-loop control of hypnosis, an RL agent was developed in the Texas Tech University Computer Science Department under the supervision of the study's principal technical investigator, Dr. Pyeatt. In cooperation with the Stanford University School of Medicine Department of Anesthesiology, Perioperative and Pain Medicine, intraoperative patient models were developed for agent training and *in silico* evaluation under the supervision of the study's principal clinical investigator, Dr. Doufas. Section 3.1 summarizes evolution of the RL agent; however, more thorough treatment of the design, development, and *in silico* testing of the RL agent may be found in our previously reported work (Moore et al., 2011a,b). After the agent was validated in simulation and the clinical study protocol gained Institutional Review Board approval, fifteen healthy volunteers underwent RL-controlled propofol hypnosis in surgical facilities of the Stanford University School of Medicine.

3.1 The Clinical-Grade RL Agent

This section addresses the development and application of the clinical-grade RL agent. As such, the architecture, training, and *in-silico* evaluation are covered. The section then presents the application of *Reagent*, a data collection and control system using the RL agent to administer propofol hypnosis in a population of healthy human volunteers.

3.1.1 Agent Architecture

The RL agent was implemented as a *Markov Decision Process* (MDP), a mathematical framework for optimal decision-making in stochastic systems. A principal feature of the MDP is the *Markov Property*, a characteristic in which the conditional probability of state transition depends solely on the action chosen in the current state—as opposed to some longer historical sequence of state visitation and action selection (Russel and Norvig, 2002; Sutton and Barto, 1998). Littman (1994) formally defines the general MDP as a system consisting of:

- the set of states $\mathcal{S} = \{s_0, s_1, s_2, \ldots, s_{|\mathcal{S}|-1}\}$,
- the transition probabilities $\Pr(s'|s,a)\,\forall s, s' \in \mathcal{S}, a \in A(s)$,
- the set of actions $\mathcal{A} = \{a_0, a_1, a_2, \ldots, a_{|\mathcal{A}|-1}\}$,
- the set of actions $A(s) \subseteq \mathcal{A}$ for each state $s \in \mathcal{S}$ that can be executed in s,
- and the immediate rewards $r^a(s)\,\forall a \in A(s), s \in \mathcal{S}$ that are available after taking any legal action from any state.

For the purposes of this research, the sets $\mathcal{S}$ and $\mathcal{A}$ were discrete. Of these components, only the transition probabilities $\Pr(s'|s,a)$ were initially unknown. (Agent training is tantamount to the discovery of these transition probabilities. Were they initially known, an optimal control policy could be determined using *dynamic programming*.) With the nature of an RL agent formally defined, discussion continues with the application of the MDP in the context of the hypnosis control task.

3.1.2 Agent Percepts

To achieve and maintain a desired level of hypnosis (BIS_{target}), the agent first observed the patient's bispectral index ($BIS_{measured}$) on five-second intervals as reported by an A-2000 BIS monitor (Covidien, Mansfield, MA). The monitor's BIS smoothing window was set to 15 seconds, the minimum, to grant the agent flexibility in managing BIS measurement noise (see Section 2.2.3).

To reduce the effect of measurement noise on the agent's estimate of patient condition, $BIS_{measured}$ was smoothed using a low-pass filter. From the resulting $BIS_{smoothed}$ signal, two control inputs were then computed: BIS_{error} and ΔBIS_{error}. BIS_{error} was defined as $(BIS_{smoothed} - BIS_{target})$, and ΔBIS_{error} was defined as the change in BIS_{error} over 15 s, or $(BIS_{error}(t) - BIS_{error}(t-2))$. These control signals allowed the agent to observe the magnitude of control error, as well its direction of change. This observation of BIS_{error} and ΔBIS_{error} served as the agent's estimation of the patient's state of hypnosis. Table 1 presents a high-level summary of patient states that may be distinguished using these control signals.

BIS_{error}	ΔBIS_{error}	Interpretation	
< 0	< 0	Good	Below target, improving
< 0	≈ 0	Neutral	Below target, steady
< 0	> 0	Poor	Below target, worsening
≈ 0	< 0	Good	At target, improving
≈ 0	≈ 0	Good	At target, steady
≈ 0	> 0	Poor	At target, worsening
> 0	< 0	Good	Above target, improving
> 0	≈ 0	Neutral	Above target, steady
> 0	> 0	Poor	Above target, worsening

Table 1: Interpreting the agent's control signals

In pilot studies of human volunteers, the combined effects of BIS measurement noise, filtering, and transport delay resulted in oscillatory control behavior. These confounding influences were successfully mitigated by conditioning BIS_{error} and ΔBIS_{error} with sets of fuzzy membership functions (Zadeh, 1965). The fuzzy set memberships for BIS_{error} and ΔBIS_{error} were assessed using two sets of triangular membership functions, $\mu_N(x)$, $\mu_Z(x)$, and $\mu_P(x)$ (Figure 1). The resulting six-dimensional feature vector served as the agent's perceptual input:

$$f = [\mu_N(E),\ \mu_Z(E),\ \mu_P(E), \mu_N(\Delta E),\ \mu_Z(\Delta E),\ \mu_P\,(\Delta E)]$$

(where E represents BIS_{error} and ΔE indicates ΔBIS_{error} for brevity). Since fuzzy set membership is expressed as a real number in the range $[0, 1]$, the continuous feature vector f required transformation before the discrete RL algorithms used in this study could be applied. Section 3.2.2 provides greater detail in the methods used to map the feature vector f to the set of discrete states $\mathcal{S}$ employed in this study.

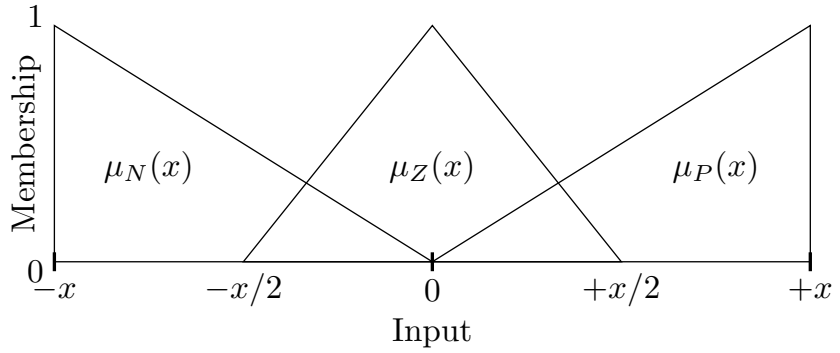


Figure 1: The system input variables, $\text{BIS}_{\text{error}}$ and $\Delta\text{BIS}_{\text{error}}$, were conditioned with sets of fuzzy membership functions. A set of three membership functions operated on $\text{BIS}_{\text{error}}$ ($x = 20$), and a second set of functions operated on $\Delta\text{BIS}_{\text{error}}$ ($x = 10$). The resulting membership values formed a six-dimensional feature vector that served as the agent's patient state descriptor.

3.1.3 Agent Actions

The agent delivered propofol to the volunteer via a catheter placed in the antecubital (elbow) vein using a precision syringe pump (Pump 33, Harvard Apparatus, Holliston, MA). During control, the agent selected an infusion rate from $\mathcal{A}$, a discrete set of 15 flowrates ranging from 0.0 – 6.0 ml/min:

$$\mathcal{A} = \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0, 6.0\} \text{ ml/min}.$$

Once a rate was selected, the chosen action remained in effect for five seconds. A concentration of 1% propofol was assumed for all members of $\mathcal{A}$.

3.1.4 Reinforcements

Although reinforcement learning is unsupervised in the sense that no explicit training exemplars are provided during training, the method assumes the existence of a *critic* that grades behavior as the agent learns. During learning, the critic's role is to dispense reinforcements in order to guide the agent's action selection. In RL, this critic is implemented as the application-specific *reward function*. In the hypnosis control task, the agent's objective was to achieve and maintain the selected BIS target. Expressed alternatively, the agent's goal was to minimize control error for the duration of the control interval. The reward function below presents one system of reinforcement for guiding action selection toward this goal

$$r(t+1) = -|BIS_{error}(t)|. \tag{1}$$

This negative-bounded reward function provided instantaneous rewards proportional to the observed control error. Under this scheme, the agent's sole means of minimizing negative reinforcement was to select actions yielding minimal control error. This reward function

highlights an important characteristic of the hypnosis control task, namely the lack of an explicit goal state. Because the task lacked a definitive persistent terminal state, it was classified as a continuing, non-episodic control task.

3.2 Agent Training

During learning, the naive, uninformed agent was expected to make arbitrarily poor propofol dosing decisions; thus, a simulated intraoperative patient was developed to facilitate agent training in a safe, off-line manner. Consequently, the principal role of this virtual patient was to model the time-dependent effects of propofol infusion, collectively known as the *pharmacokinetic* and *pharmacodynamic* (PK/PD) responses. A drug's pharmacokinetic properties describe its distribution within the body; pharmacodynamic attributes characterize the dose effect. Through experimentation with this virtual patient, the agent was expected to learn the general characteristics of propofol-induced hypnosis with respect to bispectral index: BIS is linked to propofol infusion in an inverse, time-delayed, and nonlinear manner. It should also be noted that this *in silico* patient presented an advantage in its rapid simulation of hypnotic episodes. Reinforcement learning is inherently a process of statistical estimation, and a large number of training episodes were needed to learn the control policy and achieve clinical readiness.

3.2.1 Modeling Propofol Effect

Propofol pharmacokinetics were simulated using a three-compartment model (Schnider et al., 1998), a system which uses central, rapid, and slow compartments to estimate the time-dependent distribution of propofol within the human body. In this model, propofol is introduced into the central compartment via intravenous infusion; the drug is then free to interact with the rapid and slow compartments through first-order, gradient-driven transport. These compartments, which represent collections of tissues with high and low propofol transport coefficients, were derived from empirical observations and sometimes lack direct, obvious mapping to actual physiological systems.

Figure 2 illustrates the Schnider model and its transport coefficients, which vary with patient height, weight, gender, and age. As shown, the coefficients are subscripted to indicate direction of flow (from, to) since the coefficients may differ directionally, that is, the central-to-slow coefficient (k_{cs}) differs from the slow-to-central coefficient (k_{sc}). Metabolic losses of propofol are represented in k_{c0}, establishing the only means of absolutely reducing propofol concentration. This limitation presented a substantial challenge, the agent was required to learn that inaction (realized as a zero propofol infusion rate) was the only means of decreasing hypnosis and increasing BIS.

In prior clinical study, an infusion of propofol averaged a 2.7-minute time-to-peak effect in BIS (Schnider et al., 1998). Accordingly, our PK model was augmented with a fourth *effect site* compartment. The resulting transport coefficient, $k_{e0} = 0.17$, accounted for the delay between infusion and BIS effect, which included physiological delay (mixing, circulatory, etc.) and BIS measurement delay (Doufas et al., 2004). The effect-site compartment was assumed to possess negligible volume when modeling propofol distribution.

To model the hypnotic effect of propofol, a nonlinear pharmacodynamic model was developed using previously obtained data (Doufas et al., 2004). A three-layer perceptron

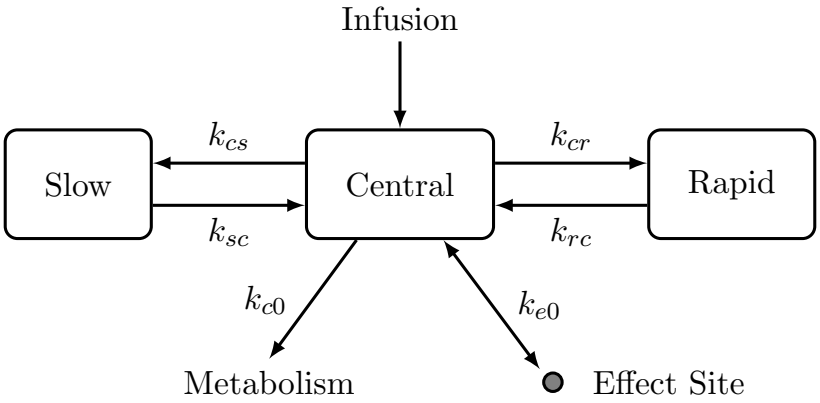


Figure 2: Schnider's pharmacokinetic model of propofol is based on the three-compartment mammalian model of pharmacokinetic action. Propofol is infused into the central volume through intravenous infusion. Concentration gradients then drive transport to the rapid and slow compartments, so named for their relative uptake rates. The site of propofol effect is modeled as an additional "virtual compartment" of infinitesimal volume in order to model observed delays between infusion and hypnotic effect.

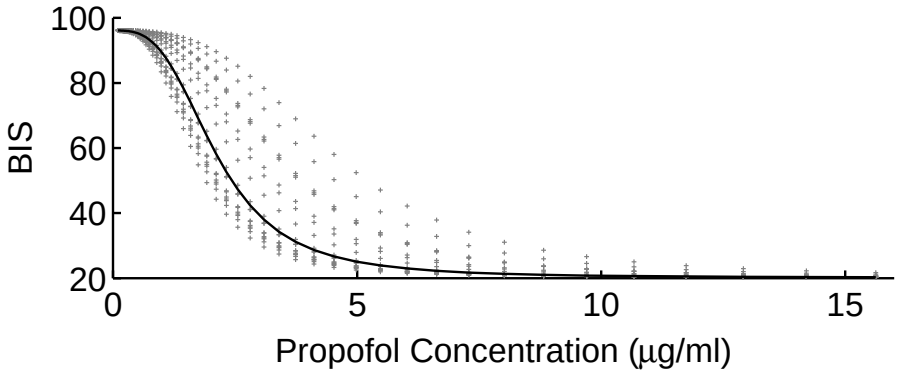


Figure 3: Doufas et al. observed the propofol/BIS response in eighteen young, healthy volunteers (2004). To model propofol pharmacodynamic effect for this study, a neural network function approximator was used to fit the median dose curve (highlighted here).

network was trained to associate arterial concentrations of propofol with observed BIS, thereby allowing the model to generally predict propofol effect from estimated effect site concentration. Figure 3 illustrates the observations of BIS and propofol concentration, as well as the median fit approximated by the neural network. The nonlinear relationship of propofol effect-site concentration to BIS is evident.

3.2.2 Knowledge Representation

During control, the RL agent is expected to observe the patient's state and then select the appropriate propofol dose using its control policy. To learn that optimal control policy, the agent accumulated its experience in *value functions*, mathematical descriptions of state utility commonly denoted as $V(s) \forall s \in \mathcal{S}$. Knowledge of V alone is not sufficient for optimal decision-making since this function only expresses the utility of an observed state. For control, it is also necessary to identify an action choice that can move the patient to more favorable conditions (or preserve existing favorable ones). In RL, the state-action value function, $Q(s,a) \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$, provides the necessary information for rational action selection. With Q, an RL agent can assess the utility of a patient state and then identify the proper infusion rate to achieve optimality for that state.

Initially, Q is unknown. The discovery of Q (learning) is accomplished through iterative function approximation. Consequently, Q must be represented in a form suitable for computational inspection and adjustment. Tables, decision trees, neural networks, and weighted polynomials have been used for this purpose in the literature. Of these, the uniformly discretized table is favored for its ease of implementation and mathematical robustness (Boyan and Moore, 1995; Baird, 1995).

In this study, state value (Q) functions were represented in a six-dimensional table. The agent's percepts, represented by the feature vector f, were mapped to a finite set of states $\mathcal{S}$ through uniform discretization. Recall that f consisted of six fuzzy state membership values (real numbers in the range $[0,1]$). To obtain a state observation $\mathcal{S}_i$ for a feature f_i, each dimension of the feature vector was partitioned into ten uniformly distributed bins, yielding a value function approximator with 10^6 entries. To permit identification of the optimal propofol infusion rate for all possible patient states, one such tabular function approximator was associated with each of the agent's actions.

3.2.3 Learning Algorithm

Watkins' Q-learning algorithm, a temporal-differencing learning method characterized by model-free, off-policy learning, was used to train the agent (Watkins, 1989). Q-learning is mathematically robust (Tsitsiklas and Van Roy, 1996; Dayan, 1992), and this robustness has contributed to the method's popularity in applied reinforcement learning. To accelerate learning, an improved form of Q-learning, called Q(λ), was applied in this study. This algorithm assimilates experience more quickly through extended temporal credit assignment. Whereas the one-step version considers only the previous action step when updating $Q(s_t, a_t)$, the improved version credits an historical chain of action selections. The "length" of this chain is governed by λ, which was set to the recommended value of 0.8 (Sutton and Barto, 1998).

3.2.4 Control Policy Identification

Watkins' Q-learning does not directly yield an optimal control policy. The algorithm only develops an approximation of the state-action value function, Q. However, the optimal control policy is trivial to determine once Q has been discovered. For each patient state s,

the optimal action choice $a^\star(s)$ may be expressed as:

$$a^\star(s) = \underset{a}{\operatorname{argmax}}\, Q\,(s,a)\ \forall a \in \mathcal{S}.$$

As a result, a state's optimal action may be represented as an integer number that indexes the ordered set of actions $\mathcal{A}$. Since the control policy identifies the best action choice for all patient states, the complete control policy may be represented as a six-dimensional table of these indices.

3.2.5 Training

Agent training consisted of a sequence of simulated hypnosis episodes using a standardized intraoperative patient prototype (male, 21 yr, 170 cm, 75 kg). To aid in learning a general association of propofol infusion and patient response, the patient's k_{e0} was randomly selected [$0.17 \pm 25\%$] at the beginning of each episode. This perturbation, of which the agent remained unaware, influenced the timing and magnitude of peak BIS effect.

To ensure sufficient exploration of the state-action space, each episode began with an *exploring start* in which a BIS target was randomly selected, and random propofol quantities were assigned to the three major PK compartments (Section 3.2.1). The agent was then permitted to interact with the patient and accumulate reinforcements for 1,000 consecutive action choices (5,000 simulated seconds). At the conclusion of an episode, a new one began with a newly randomized patient state.

Training began with a step-size parameter $\alpha = 0.2$, horizon parameter $\gamma = 0.69$, and an exploration parameter $\epsilon = 0.01$.[4] To assess the progress of learning, the sum of squared difference (SSD) was computed between intermediate control polices. When the SSD metric fell below a small threshold θ, α was halved, and learning resumed. This procedure continued until α measured 10^{-5} or less. In total, training required 5×10^7 episodes over approximately one week of CPU time on a contemporary desktop computer.

3.3 *In silico* Control Policy Evaluation

Prior to clinical application, the agent was evaluated in simulation to assess the fitness of the agent and its control policy. Although the agent was trained using an ideal simulated patient (fixed demographic parameters and near-ideal PK/PD characteristics), an actual surgical patient was not expected to present so favorably. Because intraoperative patients vary in height, weight, age, and gender (and other attributes), their PK/PD responses to propofol cannot be so neatly characterized.

To challenge the agent in a more realistic manner, the patient model illustrated in Figure 4 was modified to express patient-specific variation. The first point of individual variability was found in simple demographics. Schnider reported lean body mass, age, and gender to be significant covariates in propofol pharmacokinetic response (Schnider et al., 1998). Accordingly, the RL agent was tested on simulated patients with a range of demographic parameters. Since the Schnider model considers these parameters in its estimation of propofol distribution, demographic variation was not judged sufficient challenge for the agent.

4. For a more thorough discussion of these parameters and their import, see Moore et al. (2011b).

For additional challenge, the ideal PK and PD models were perturbed in ways mimicking the variation routinely observed in the operating room. A few quantitative (Röpcke et al., 2001b; Schnider et al., 1999, 1998; Bailey et al., 1996) and qualitative (Kearse Jr. et al., 1994; Ausems et al., 1986) descriptions of intraoperative patient variability may be found in the anesthesia literature. Taken collectively, evidence suggests that individual patient variation may be expressed as deviation in propofol pharmacokinetics (Gentilini et al., 2000; Schwilden et al., 1987) or pharmacodynamics (Struys et al., 2004, 2001).

In this study, individualized response in the simulated intraoperative patient was achieved with a *Patient Variability Model* (PVM), a mechanism for perturbing the patient's PK/PD in a manner removed from the agent's direct observation (illustrated in Figure 7). The PVM was implemented as two distinct components: one which affected the ideal pharmacokinetics (PK_{PVM}), and one which perturbed ideal pharmacodynamics (PD_{PVM}).

3.3.1 Pharmacokinetic Variation

The anesthesia literature provides evidence that patients commonly exhibit pharmacokinetic variation. Gepts observed: "When individuals are given identical doses per kg of body weight, large differences in pharmacological response may be seen" (Gepts, 1998, 10) and "Pharmacokinetic variability is much greater in sick compared with healthy people..." (Gepts, 1998, 11). The findings of Doufas et al. support those observations of variability. In a propofol pharmacokinetic study of 18 healthy volunteers, k_{e0} was determined to be 0.17 min^{-1} (range $[0.08, 0.25]$ min^{-1}) (Doufas et al., 2004). To model this source of patient variation, the PK_{PVM} block varied k_{e0} in conjunction with variation in patient demographics.

Figure 5 illustrates the effect of k_{e0} variation in simulated patients. A bolus of propofol was applied at $t = 0$ min and allowed to distribute under the Schnider pharmacokinetic model at the selected k_{e0} values. As shown, larger k_{e0} coefficients represented more "tightly coupled" systems in which propofol was transported to the effect site more readily, resulting in deeper hypnosis for a given dose. For emphasis, Figure 5 highlights the minimum hypnotic levels, as well as the times of their occurrence. While the time of peak effect varied by approximately 25 seconds, the range in peak effect varied by more than 20 points, a range that can span the clinically meaningful endpoints of light to deep hypnosis (as measured by BIS).

3.3.2 Pharmacodynamic Variation

Other sources of patient variation were better modeled as perturbations in propofol pharmacodynamics (i.e., effect, rather than distribution). For example, propofol sensitivity or tolerance may be modeled intuitively as a respective heightened or attenuated pharmacodynamic response to a given concentration of propofol. Exogenous factors, such as measurement noise and surgical stimuli, can not be reasonably expected to alter the pharmacokinetic distribution of propofol within the patient; however, these influences may directly alter the hypnotic action of the drug.

The role of the PD_{PVM} block was to model those factors best expressed as change in pharmacodynamics. The PD_{PVM} block accomplished this by decomposing pharmacodynamic variability into three classes: propofol sensitivity, intraoperative stimuli, and measurement

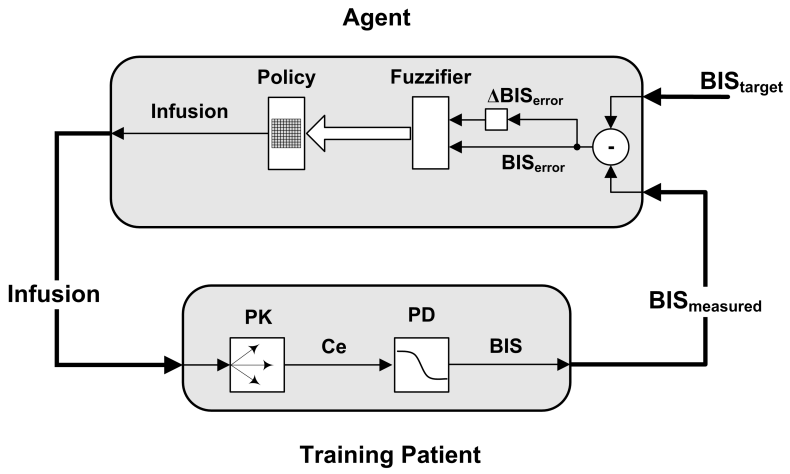


Figure 4: This figure illustrates the agent and its relationship to the simulated intraoperative patient used for training. The agent observed two external inputs, $\mathrm{BIS}_{\mathrm{target}}$ and $\mathrm{BIS}_{\mathrm{measured}}$, to compute the control error ($\mathrm{BIS}_{\mathrm{error}}$) and the change in control error over time ($\Delta\mathrm{BIS}_{\mathrm{error}}$). The intraoperative patient was modeled with near-ideal propofol PK/PD parameters.

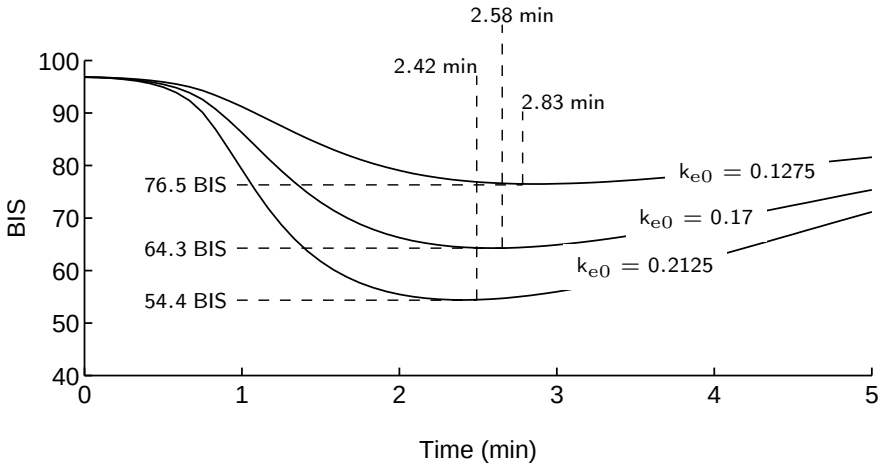


Figure 5: To demonstrate the variation associated with changes in k_{e0}, a bolus of propofol was delivered to a simulated patient, and distribution of propofol was modeled over time. For comparison, k_{e0} was selected at values of 0.17, 0.1275 (0.17-25%), and 0.2125 (0.17+25%). The points of peak BIS effect and their associated times are highlighted.

noise. The simulated patient's ideal BIS was then perturbed with a sum of time-dependent and independent combinations of these factors.

Modeling changes in propofol sensitivity begins as a straightforward process. In the tolerant patient, a given concentration of propofol may produce higher-than-expected BIS levels (more conscious than predicted). Conversely, lower-than-expected BIS levels may be observed in the patient with increased propofol sensitivity (less aware than predicted). In the evaluation of the RL agent, these differences were considered as a constant bias in pharmacodynamic effect. This time-independent parameter, denoted as $\Delta\text{BIS}^{\text{i}}_{\text{static}}$, was implemented as an additive factor, ≥ 0 in the resistant patient and ≤ 0 in the sensitive patient.

A credible model of propofol sensitivity should also consider exogenous sources of variation, one of which is noxious surgical stimuli. When studying patient variation, it is reasonable to conclude that some surgical procedures are more painful than others. For example, in a common heart procedure, such as the coronary artery bypass graft (CABG), the patient's chest is opened with an approximate six-inch incision, and the sternum is separated for access to the heart. Compare this procedure to the arthroscopic repair of a rotator cuff injury and its small 1-cm incisions. Intuitively, the degree of noxious stimulation in the CABG procedure is expected to exceed that of rotator cuff repair. Ausems et al. support this expectation in a report that found upper abdominal procedures required more analgesia than other smaller procedures (Ausems et al., 1986). Likewise, intraoperative stimuli were correlated to increased analgesic requirements in patients undergoing lower abdominal gynecologic, upper abdominal, and breast surgery. From these observations, as well others (Barvais et al., 1996), it is reasonable to conclude that some surgical procedures are inherently more noxious than others. Accordingly, the time-independent positive constant $\Delta\text{BIS}^{\text{s}}_{\text{static}}$ was used to denote this persistent noxious surgical stimulus.

Noxious stimulus may also be presented in a time-dependent manner. Absalom observes, "It is not always possible to predict when a surgeon will suddenly inflict a noxious stimulus on the patient..." (Absalom et al., 2002, 73). Ausems et al. reported that different intraoperative stimuli, including tracheal intubation, skin incision, and closure, required different levels of analgesia to maintain satisfactory anesthesia (Ausems et al., 1986). More recently, decreases in hypnotic level have been associated with surgical stimulation (Röpcke et al., 2001b), while increases in bispectral index have been correlated with skin incision (Kearse Jr. et al., 1994). Ausems et al. also observed that "single short-duration" stimuli, such as skin incision, required higher concentrations of opioid analgesic to ensure adequate anesthesia (Ausems et al., 1986). Given these observations, the short-duration surgical stimulus can reasonably be considered a transient perturbation in propofol pharmacodynamics that presents as a temporary decrease in hypnosis.

The effects of intraoperative stimuli are not limited to arousal events, those that decrease hypnotic effect. Röpcke found that concomitant administration of propofol and remifentanil (an opioid analgesic) resulted in lower than expected measurements of bispectral index in the intraoperative patient (Röpcke et al., 2001a). Whereas the noxious stimulus could be viewed as transient propofol tolerance, this synergistic drug interaction may present as temporarily heightened propofol sensitivity. These depressive events pose a unique challenge for hypnosis control since the agent cannot directly intervene and reduce the patient's propofol concentration.

During *in silicon* verification of the RL agent, irregular transient stimuli were presented to the agent to evaluate its ability to handle the dynamic conditions commonly found in the

intraoperative patient. To challenge the agent in an unpredictable manner, the duration, timing, and intensity of the short-duration stimuli were randomly chosen. In addition, the "direction" of challenge was randomized. A positive magnitude, which simulated an arousing event, was chosen with probability 0.8. Depressive events were chosen with probability 0.2. For this study, the time-dependent affect on patient pharmacodynamics was denoted as $\Delta\text{BIS}_{\text{dynamic}}(\text{t})$, where t indicated time dependence.

Finally, BIS is intrinsically noisy since the underlying EEG is a low-power signal requiring amplification for adequate measurement (as discussed previously). Prior study has modeled this noise as a stationary, normally-distributed signal ($\mu = 0$, $\sigma = 3$) (Struys et al., 2004). We modeled BIS measurement noise in accordance with this precedent.

In summary, the PVM modeled individual patient variability with changes in propofol pharmacokinetics and pharmacodynamics that remained hidden from agent observation. The PK_{PVM} component modeled changes in k_{e0}, while the PD_{PVM} block modeled changes in propofol sensitivity ($\Delta\text{BIS}_{\text{PVM}}$) as a sum of time-dependent and time-independent parameters (Moore et al., 2009). The cumulative PVM influence can thus be summarized as:

$$\Delta BIS_{static} = \Delta BIS^{i}_{static} + \Delta BIS^{s}_{static},$$

$$\Delta BIS_{PVM}(t) = \Delta BIS_{static} + \Delta BIS_{dynamic}(t) + \mathcal{N}(0,3),$$

$$BIS_{measured}(t) = BIS_{ideal}(t) + \Delta BIS_{PVM}(t).$$

3.4 Assessment of Agent Performance

The clinical study protocol included performance analysis of RL control under steady-state (maintenance of hypnosis) and non-steady-state conditions (induction of hypnosis and change in $\text{BIS}_{\text{target}}$). In the clinical practice of anesthesia, precise control has less value during non-steady-state periods. Conditions that a control engineer might consider unfavorable, like target overshoot, are expected during a manual induction as the clinician seeks to quickly achieve the desired target. Because the agent's principal goal of fine control is less relevant during induction, performance analysis of this interval has been omitted from this discussion.

3.4.1 Evaluation Population

To evaluate the agent's ability to provide well-controlled propofol hypnosis in a diverse population, a set of 1,000 individualized patients was generated *in silico*. Control performance was assessed over one episode of hypnosis for each of these individualized patients. In each episode, the agent was first presented a fully conscious patient and then tasked with achieving and maintaining propofol-induced hypnosis for 240 minutes, an interval that the clinical team considered representative. During the episode, BIS targets were randomly selected (without replacement) from the set {40,50,60}. Once selected, a target remained in effect for 80 minutes.

3.4.2 Maintenance Interval Identification

Although the induction interval is not addressed in this discussion, induction, along with BIS target change events, delimit the maintenance control intervals. The first maintenance

interval began with the completion of anesthetic induction. The induction period began when RL control was engaged to induce anesthesia in the conscious simulated patient (BIS $\approx$ 95). Induction continued until steady-state conditions, as defined by O'Hara et al., were observed at the selected target (1992). (The O'Hara metrics include T_{sp} = Time to Setpoint,T_{peak} = Time of Peak BIS, T_{sp} = Time to Steady State, and BIS_{peak} =$BIS_{measured}$ at T_{peak}.) This steady-state point, identified as T_{ss} in Figure 6, marked the beginning of the first maintenance control interval. The maintenance interval continued until the time of BIS target change, denoted as $T_{ss} + 80$ min, or ΔBIS_{target}.

The beginning of the next maintenance period was delineated similarly since the conditions at target transition resembled those at induction. After a step change in BIS_{target}, the agent acted to reestablish control and achieve steady-state conditions at the new target. Accurate identification of the new steady state was slightly complicated. High-to-low target changes (i.e., BIS_{target}=60 to BIS_{target}=40) directly compared to induction, while low-to-high changes (i.e., BIS_{target}=40 to BIS_{target}=50) resembled induction in an *inverted* sense. Once the new T_{ss} was achieved, the second maintenance control interval continued until the second target change.

The beginning of the third maintenance period was handled just as the second maintenance period. However, this control period was terminated by the end of automated control. Propofol infusion was discontinued, and the virtual patient was allowed to recover normal consciousness.

3.4.3 Performance Metrics

The steady-state control performance was evaluated using the four metrics of Varvel et al. (1992), which comprise the standard performance measures in closed-loop infusion control. These metrics build upon the instantaneous performance error (PE):

$$PE = \frac{BIS_{smoothed} - BIS_{target}}{BIS_{target}} \cdot 100. \tag{2}$$

The first metric, the median performance error (MDPE), indicates the control bias observed in a single patient and is computed as

$$MDPE_i = \text{median}\,(PE_{ij}) \qquad j = 1 \ldots N, \tag{3}$$

where i identifies a subject, and j iterates over the set of PE measurements for a subject. Median absolute performance (MDAPE) error reflects the accuracy of the controller in a subject:

$$MDAPE_i = \text{median}\,(|PE_{ij}|) \qquad j = 1 \ldots N. \tag{4}$$

Wobble measures the intra-subject variability in performance error:

$$Wobble_i = \text{median}\,(|PE_{ij} - MDPE_i|)\ j = 1 \ldots N. \tag{5}$$

Divergence is defined as the slope of the regression line computed through the observed MDAPE measurements. Positive values indicate an increasing difference in measured and target values; a negative divergence indicates more stable control.

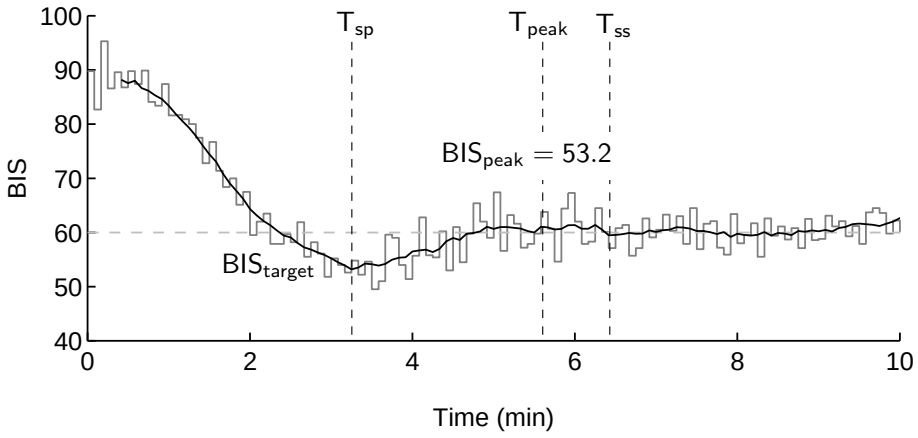


Figure 6: During analysis of control performance, the dynamic performance parameters T_{sp},T_{ss}, T_{peak}, and BIS_{peak} first reported by O'Hara et al. (1992) were programmatically identified to precisely delineate the maintenance control periods.

In addition to the Varvel metrics, contemporary studies of closed-loop anesthesia report the *Controlled* metric, the percentage of measurements in which the measured BIS was observed to be within ± 10 BIS (Struys et al., 2004) or ± 5 BIS (De Smet et al., 2008) of target. As an additional performance comparator, this study also reports the root-mean-square error (RMSE) computed for each maintenance control interval.

3.4.4 Acceptance Criteria

The anesthesia literature does not provide a definitive guideline for clinically suitable control of propofol-induced hypnosis, but a survey of three contemporary studies (De Smet et al., 2008; Struys et al., 2004; Absalom and Kenny, 2003) provides some reasonable performance goals (Table 2). These performance objectives should be interpreted carefully since specific values of these measures have not been correlated to favorable clinical outcomes. In other words, no study strongly indicates that an MDPE of 5% is x times better than an MDPE of 10%. In the absence of such data, we aimed for performances levels that surpassed reported values by reasonable margins.

3.4.5 Simulation Results

As indicated by comparison of observed performance and respective targets (Tables 2 and 3), the median values for all observed steady state parameters met their respective acceptance criteria. MDPE, MDAPE, Wobble, Divergence, and RMSE all presented values below the respective requirements. (Note the change in units in the Divergence measure.) Likewise, the Controlled metric was above its minimum threshold. These results suggested that the RL agent was suitable for evaluation in healthy volunteers. However, no definitive conclusion could be drawn since the accuracy of the Patient Variability Model was not verified prior

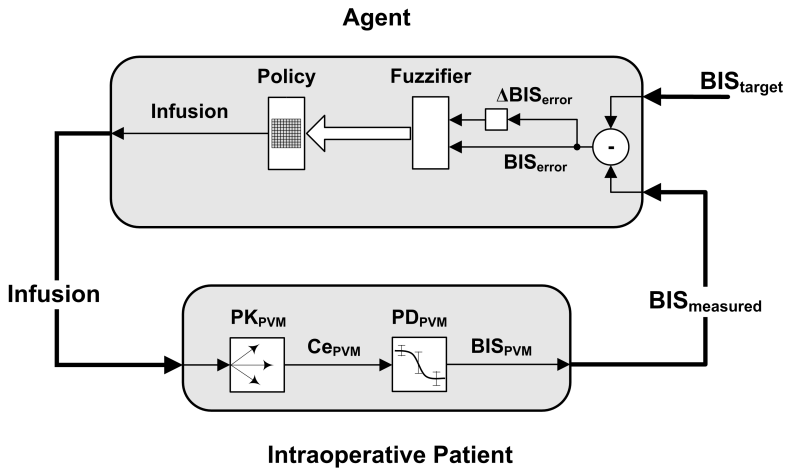


Figure 7: This figure illustrates the agent and its relationship to the simulated intraoperative patient used for evaluation. Like the training system, the agent relied on $\mathrm{BIS}_{\mathrm{target}}$ and $\mathrm{BIS}_{\mathrm{measured}}$ to compute control error, as well as change in control error. Unlike the training system, the intraoperative patient presented variable propofol PK/PD responses.

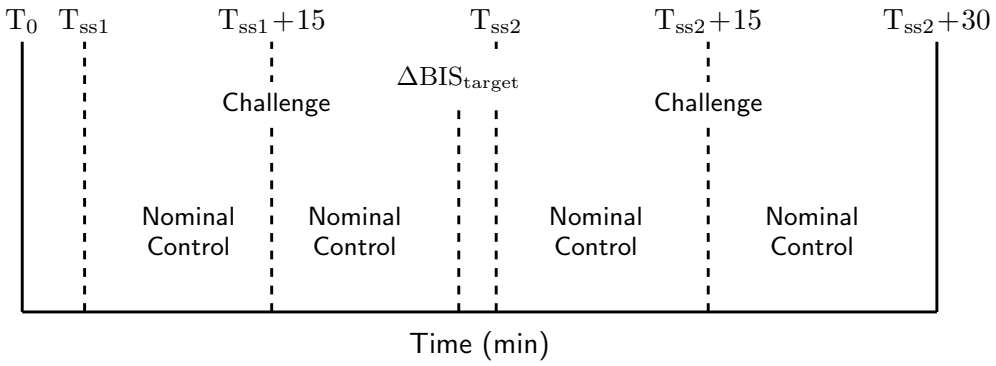


Figure 8: Induction (T_0) marked the beginning of the first BIS target evaluation period. Nominal control periods, as well as the surgical challenge, were scheduled in relation to $\mathrm{T}_{\mathrm{ss1}}$ (the time at which steady-state control was observed). Control continued to the maintenance interval's end at ($\mathrm{T}_{\mathrm{ss1}}$ + 30) min. At that time, a new $\mathrm{BIS}_{\mathrm{target}}$ was selected (labeled $\Delta\mathrm{BIS}_{\mathrm{target}}$ here) and a second, similar event schedule was observed. Recovery began at ($\mathrm{T}_{\mathrm{ss2}}$ + 30) min after automated control was discontinued.

to this study. After review of the simulation protocol and results, the principal clinical investigator granted approval for human study.

Parameter	Criterion
MDPE‡	± 5.0
MDAPE‡	7.5
Wobble‡	5.0
Divergence⋆	± 0.001
Controlled‡◇	80
RMSE§	5.0

‡(%), ⋆(%/hr), §(BIS)

◇ Percentage of time within ±5 BIS of target.

Table 2: Steady state performance acceptance criteria

Parameter	Observation	
MDPE‡	-0.17	(-0.50, 0.25)
MDAPE‡	3.33	(3.17, 3.50)
Wobble‡	3.30	(3.13, 3.50)
Divergence⋆	0.001	(-0.001, 0.003)
RMSE§	2.79	(2.58, 3.07)
Controlled‡◇	82.4	(80.6, 84.0)

median (IQR) ‡(%), ⋆(%/hr), §(BIS)

◇ Percentage of time within ±5 BIS of target.

Table 3: Simulated steady-state performance metrics

3.4.6 Clinical Application of RL Control

After IRB approval in the Stanford University School of Medicine, we recruited fifteen healthy (BMI $\leq$ 25 kg/m^2, 18–45 yr) volunteers. The clinical study was conducted in an operating room in the Stanford University Medical Center under informed consent. To facilitate clinical study, a custom data collection and control system, dubbed *Reagent*, was developed. The hypnosis control hardware consisted of a standard desktop computer, an A-2000 BIS monitor (Covidien, Mansfield, MA), and a Harvard Pump 33 dual syringe pump (Harvard Apparatus, Holliston, MA). The software consisted of a graphical user interface for clinician use, an embedded RL agent for propofol dosing, and various other modules for BIS monitor and syringe pump communication.

Volunteers fasted for at least six hours prior to the study and their vital signs were monitored according to the standards of the American Society of Anesthesiologists (ASA). After placement of the monitors, an intravenous catheter was inserted at the elbow for agent-directed propofol infusion. The study began when the anesthesiologist engaged RL control to achieve a randomly selected initial target (40 or 60). Once $\text{BIS}_{\text{target}}$ was achieved, the agent was permitted to regulate the level of hypnosis undisturbed for 15 minutes. A tetanic stimulus was then administered to the volunteer's thigh to simulate a noxious, destabilizing surgical event. Control was allowed to continue for an additional 15 minutes

(see Figure 8). At that time, the agent was directed to achieve the second BIS_{target}. Once the volunteer had stabilized at the second target, a similar procedure of maintenance and stimulus followed. Finally, automated hypnosis control was disengaged, and the volunteer was allowed to recover normal consciousness.

3.4.7 Performance Analysis

The agent's steady state control performance was assessed using the same procedures applied in the *in silico* evaluation. Automated tools identified induction, maintenance, and target change intervals. Maintenance intervals were then scored using the methods applied in the *in silico* performance analysis (Equations 2–5). The $BIS_{target} = 40$ and $BIS_{target} = 60$ control periods were evaluated independently and then in aggregate form.

The expected infrequency of BIS target change and relatively short duration between targets place the importance of transition control performance below maintenance performance; however, well-controlled behavior during BIS target change remains valued since the patient's need for hypnosis may vary over the course of the surgical procedure. In response, the dynamic O'Hara metrics are also presented in order to more thoroughly characterize the agent's control abilities. As discussed previously, these metrics were programmatically determined to identify maintenance intervals and were readily available.

4. Results

This section tabulates the study's observations. Results were grouped into three primary sets for analysis: Target 40 Maintenance, Target 60 Maintenance, and Aggregate Maintenance. The subordinate transition control metrics are reported, as well.

4.1 Volunteers

Fifteen healthy volunteers (11 males and 4 females) were recruited for the study of agent-guided propofol hypnosis. Table 4 presents the observed demographic parameters, and Table 5 summarizes those parameters. As the tables show, the volunteer population appeared to be young, healthy (ASA I), and predominantly male—characteristics reflecting the student population with ready access to study recruitment postings.

4.2 Target 40 Maintenance Control Metrics

Figure 9 graphically illustrates the $BIS_{measured}$ and $BIS_{predicted}$ values observed in each of target 40 episodes. The X-axes have been standardized to a 30-minute window. Note that the duration of an episode did not always equal 30 minutes due to timing differences between events hand-marked during the study and the more rigorous, post-study automated segmentation. The Y-axes have been standardized to a 60-BIS interval.

Some immediate observations can be made from Figure 9. The vertical black line indicates the time at which the tetanic stimulus was applied. Volunteers 5, 6, 7, 14, and 15 showed clearly distinguished arousal responses to noxious stimuli. The figure also highlights notable behavior in the predicted BIS. In most Target 40 episodes, $BIS_{predicted}$ demonstrated marked deviation from the observed BIS ($BIS_{measured}$). The degree of mis-prediction varied with volunteer, and prediction error appeared to vary within individual volunteers in a

ID	Gender	Age (yr)	Weight (kg)	Height (cm)	BMI (kg/m^2)
01	Male	21	84.0	183	25.1
02	Male	18	63.6	173	21.2
03	Male	20	69.0	178	21.8
04	Male	20	77.0	185	22.5
05	Male	18	61.4	175	20.1
06	Female	19	50.0	152	21.5
07	Male	22	77.3	178	24.4
08	Female	26	56.8	163	21.4
09	Female	19	61.4	163	23.1
10	Male	21	75.0	188	21.2
11	Male	19	70.5	180	21.7
12	Male	25	82.0	183	24.5
13	Male	20	61.4	175	20.0
14	Male	24	60.0	173	20.1
15	Female	19	59.1	168	20.9

Table 4: Human subject demographics

Age (yr)	Weight (kg)	Height (cm)	BMI (kg/m^2)
20.7 ± 2.5	72.2 ± 10.0	174.5 ± 9.6	22.0 ± 1.6
Mean $\pm$ SD		$n = 15$ ($n_{male} = 11, n_{female} = 4$)	

Table 5: Human subject demographic summary

time-dependent manner. No obvious systematic bias is evident; the model over-predicted in some volunteers but under-predicted in others.

Table 6 presents the observed Target 40 control metrics for each volunteer. The metrics are generally indicative of good control; however, two notable exceptions appear in the table. First, Volunteer 2's control metrics stand out as outliers. In this case, the volunteer exhibited strong bouts of coughing at Target 40. Although signs of illness were not obvious prior to study, the volunteer admitted to "having a cold" in a post-study interview. Likewise, Volunteer 11's study duration is anomalous; this Target 40 interval was abbreviated due to mis-configuration of the syringe pump after a fresh syringe was loaded.

4.3 Target 60 Maintenance Control Metrics

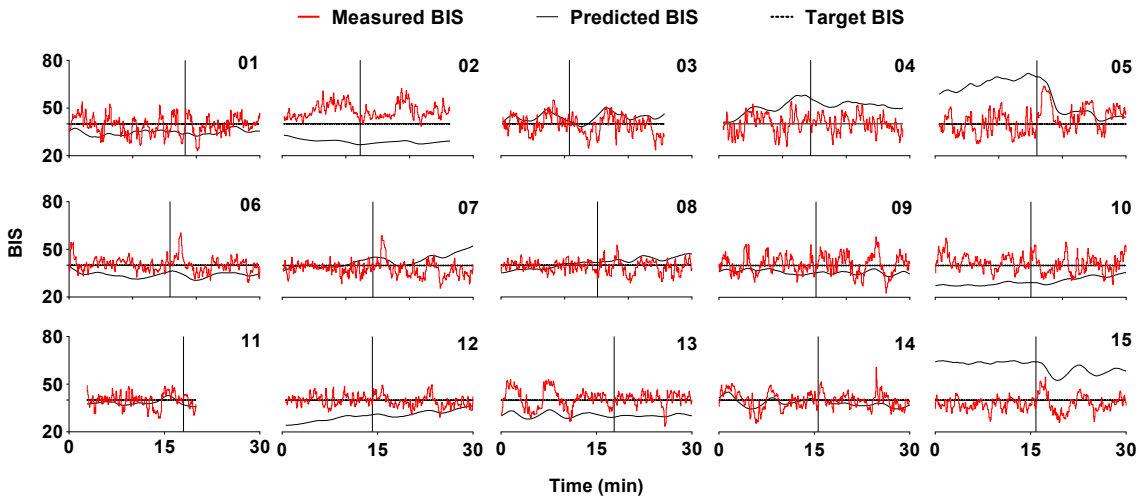
Figure 10 follows the format of Figure 9 in illustrating the BIS values observed in the Target 60 episodes. As before, the vertical black line indicates the point of tetanic stimulus. Volunteers 4, 5, 6, 7, 8, 13, 14, and 15 show responsive arousal behavior. Not unexpectedly, more volunteers exhibited obvious stimulus responses at this lighter hypnosis level. Table 7

ID	Target Order	Duration (min)	MDPE (%)	MDAPE (%)	Wobble (%)	Divergence (%/hr)	RMSE (BIS)	Controlled (%)
01	2	39.5	-3.5	11.0	10.5	-0.0002	6.5	56.8
02	1	26.1	17.2	17.2	7.2	0.0002	9.4	27.7
03	2	25.4	-0.1	11.2	11.1	0.0002	5.9	59.5
04	1	28.2	2.0	9.8	9.9	-0.0002	5.6	60.9
05	2	30.3	5.5	15.0	15.5	0.0004	8.4	42.2
06	1	36.3	1.8	6.8	7.0	-0.0001	4.9	77.6
07	2	30.2	-4.8	7.0	6.6	0.0004	5.1	69.8
08	1	30.2	-2.0	6.5	6.0	0.0002	3.9	81.3
09	2	30.8	2.0	10.2	10.5	0.0002	6.0	59.6
10	1	35.4	5.1	10.5	10.4	-0.0006	7.6	59.2
11	2	17.2	0.5	6.5	6.2	0.0004	4.5	72.0
12	1	29.6	-1.2	5.7	5.5	0.0001	3.6	84.3
13	1	32.9	2.2	8.8	7.8	-0.0005	5.4	67.2
14	1	30.6	-0.3	8.0	8.1	-0.0003	5.0	69.8
15	2	30.5	-6.0	8.8	7.0	0.0001	5.1	67.6

Table 6: Observed performance at BIS_{target}=40

ID	Target Order	Duration (min)	MDPE (%)	MDAPE (%)	Wobble (%)	Divergence (%/hr)	RMSE (BIS)	Controlled (%)
01	1	27.7	-1.0	1.4	1.5	0.0003	3.5	89.8
02	2	29.2	1.3	2.8	2.5	0.0001	3.1	90.3
03	1	24.6	0.7	3.2	3.2	0.0000	2.7	94.9
04	2	30.7	0.7	2.5	2.2	0.0000	2.2	97.8
05	1	30.6	-1.8	4.2	3.8	0.0003	4.6	77.4
06	2	32.2	0.6	4.0	3.8	0.0001	4.5	79.6
07	1	29.0	-0.7	2.8	2.7	0.0001	2.8	91.1
08	2	33.4	-0.7	4.5	4.7	0.0001	5.2	71.4
09	1	27.7	2.8	4.3	3.0	0.0000	4.0	80.8
10	2	30.0	0.2	3.5	3.5	0.0000	3.7	82.5
11	1	35.0	0.2	1.8	2.0	-0.0002	3.8	89.1
12	2	31.0	1.7	3.3	2.7	-0.0001	3.1	89.5
13	2	30.6	0.8	2.3	2.2	-0.0001	2.6	92.1
14	2	29.5	-2.3	6.7	5.7	0.0000	5.8	59.7
15	1	30.4	0.2	4.0	3.8	0.0001	4.2	79.5

Table 7: Observed performance at BIS_{target}=60



*Volunteer 11's episode was interrupted due to misconfiguration of the syringe pump.

Figure 9: The figure presents the observed BIS measurements at BIS_{target}=40. The X-axis represents a 30-minute window of time, and the Y-axis represents a 60-BIS interval. Each plot is labeled with the respective volunteer ID, and the vertical black line identifies the time of noxious stimulus. As shown, Volunteers 5, 6, 7, 14, and 15 demonstrated clear arousal responses to noxious stimulus.

presents the observed Target 60 control metrics for each volunteer. All of these metrics indicative of good hypnosis control. It should also be noted that these results are similar to those reported in the simulation (Table 3). Like the Target 40 observations, the Target 60 cases displayed varying degrees of PK/PD model mis-prediction. No systematic bias was detected.

4.4 Aggregate Maintenance Control Metrics

Table 8 presents each volunteer's aggregate control results. These metrics were computed by pooling the Target 40 and Target 60 observations to produce a set of global control measures. Table 9 summarizes the control metrics of the three groups (Target 40, Target 60, and Aggregate) with basic descriptive statistics. As shown, the mean aggregate control metrics exceed the desired performance levels presented in Table 2. The individual results were mixed: performance at BIS_{target}=60 met the desired goals by comfortable margins, while performance at BIS_{target}=40 narrowly missed desired levels in the Controlled and Wobble metrics.

4.5 Target Transition Metrics

Table 10 presents the observations obtained at changes from BIS Target 60 to Target 40. This high-to-low target change presented the most direct transition, and these observations were consistent with those observed in simulation. Table 11 presents the observations

ID	Duration (min)	MDPE (%)	MDAPE (%)	Wobble (%)	Divergence (%/hr)	RMSE (BIS)	Controlled (%)
01	67.2	-2.5	7.0	6.8	0.0000	5.3	70.5
02	55.3	8.8	9.6	4.7	0.0001	6.1	60.8
03	50.0	0.3	7.3	7.2	0.0001	4.3	76.9
04	59.0	1.3	6.0	5.9	-0.0001	3.8	80.1
05	60.9	1.8	9.6	9.6	0.0004	6.5	59.9
06	68.6	1.2	5.5	5.5	-0.0000	4.8	78.5
07	59.2	-2.8	5.0	4.7	0.0003	4.0	80.2
08	63.6	-1.3	5.4	5.3	0.0002	4.6	76.1
09	58.6	2.4	7.4	6.9	0.0001	5.1	69.6
10	65.4	2.9	7.3	7.2	-0.0003	5.8	69.9
11	52.2	0.3	3.4	3.4	-0.0000	4.0	83.4
12	60.6	0.2	4.5	4.1	0.0000	3.3	87.0
13	63.5	1.6	5.7	5.1	-0.0003	4.1	79.2
14	60.1	-1.3	7.3	6.9	-0.0002	5.4	64.9
15	60.9	-2.9	6.4	5.4	0.0001	4.6	73.5

Table 8: Observed aggregate maintenance performance

	BIS_{target} 40	BIS_{target} 60	Aggregate
Duration†	30.2 ± 5.2	30.1 ± 2.5	60.3 ± 5.1
MDPE‡	1.0 ± 5.6	-0.2 ± 1.2	0.4 ± 3.0
MDAPE‡	7.4 ± 3.5	2.8 ± 1.2	5.1 ± 1.7
Wobble‡	6.2 ± 2.6	2.6 ± 1.2	4.5 ± 1.5
Divergence*	< 0.001	< 0.001	< 0.001
RMSE§	4.5 ± 1.7	2.9 ± 1.1	3.7 ± 0.9
Controlled‡	79.0 (70.9, 89.0)	92.8 (83.3, 100.0)	85.5 (72.9, 88.5)
Mean ± SD Median (IQR)	n=15	†(min), ‡(%), *(%/hr), §(BIS)	

Table 9: Summary of observed maintenance performance

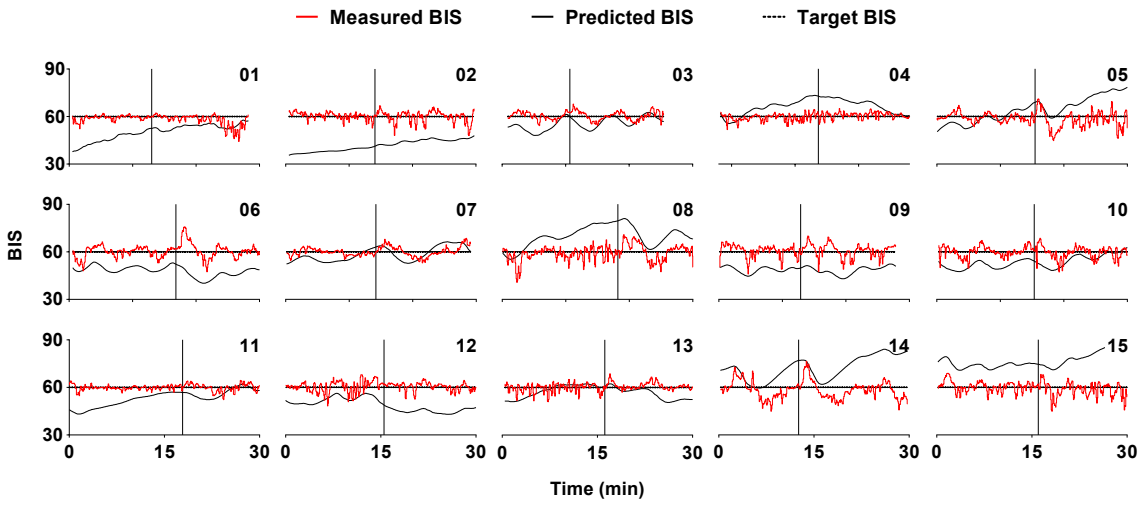


Figure 10: The figure presents the observed BIS measurements at BIS_{target}=60. The X-axis represents a 30-minute window of time, and the Y-axis represents a 60-BIS interval. Each plot is labeled with the respective volunteer ID, and the vertical black line identifies the time of noxious stimulus. As shown, Volunteers 3, 5, 6, 7, 8, 13, 14, and 15 demonstrated clear arousal responses to noxious stimulus.

Vol	T_{sp} (min)	T_{peak} (min)	T_{ss} (min)	BIS_{peak} (BIS)
1	2.04	4.62	7.62	11.13
3	2.46	4.21	4.79	6.49
5	1.47	2.30	3.13	14.87
7	2.84	4.42	5.92	5.13
9	2.61	3.61	4.44	10.88
11	3.24	3.74	4.24	3.14
15	2.67	4.76	6.34	6.12
	2.47 ± 0.57	3.95 ± 0.84	5.21 ± 1.51	8.25 ± 4.13
Mean ± SD				n=7

Table 10: Observed transition performance: Target 60 to 40

associated with changes from BIS Target 40 to Target 60. All three time measurements exceed those high-to-low transition observations by notable margins. These observations may be initially interpreted as evidence of the previously-discussed asymmetrical control influence.

Vol	T_{sp} (min)	T_{peak} (min)	T_{ss} (min)	BIS_{peak} (BIS)
2	7.13	8.13	8.13	1.29
4	11.25	12.09	14.25	14.22
6	11.19	11.94	16.61	13.06
8	11.39	12.36	14.86	15.91
10	10.79	12.54	14.20	13.66
12	10.84	12.17	14.26	10.36
13	6.58	7.58	7.58	2.10
14	8.59	8.92	9.59	0.38
	9.72 ± 1.99	10.71 ± 2.11	12.43 ± 3.45	8.87 ± 6.51

Mean ± SD n=8

Table 11: Observed transition performance: Target 40 to 60

5. Discussion

This section presents additional discussion highlighting the promising aspects of RL in closed-loop anesthesia control. The section also identifies some limitations of the clinical study and presents some opportunities for future research.

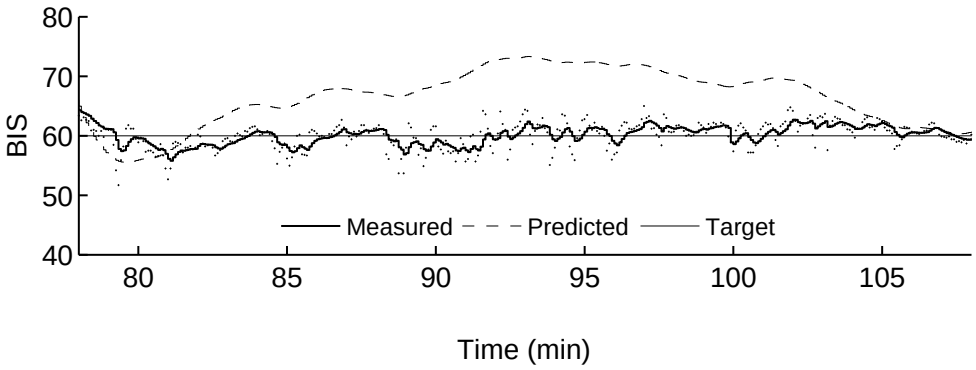
5.1 Clinically-acceptable Performance

The RL agent delivered propofol hypnosis in a manner consistent with well-controlled anesthesia, and control performance met or exceeded most performance targets. Control was considered accurate, as measured by MDPE, RMSE, and Controlled Percentage. The negligible Divergence values indicated that control was stable. The MDAPE and Wobble metrics were generally good, although an undesirable degree of oscillation was observed in some volunteers.

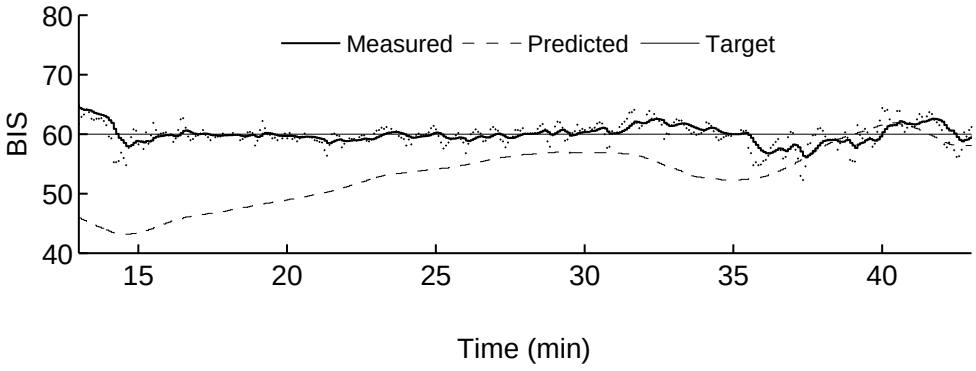
Furthermore, the agent demonstrated resistance to the disrupting tetanic stimulus. In cases where clear arousal response was observed, the agent reasserted control after the noxious event (Figures 9 and 10). Testing the boundaries of agent's capabilities in this manner is enlightening, but may be overly aggressive because propofol does not provide analgesia and cannot effectively manage pain as well as other anesthetics. In the intraoperative setting, propofol is commonly administered alongside an opioid analgesic, drugs that tend to suppress pain-induced arousal events, like those observed in this study.

5.2 Patient-specific Hypnosis

Figure 11 illustrates one favorable aspect of RL control: patient-specific hypnosis. During each study, the data collection system computed the predicted bispectral index as the agent controlled the volunteer's level of hypnosis. Using the volunteer's demographic data, the agent's action history, and the Schnider-Doufas PK/PD model, an estimate of propofol effect was computed on five-second intervals. By comparing predicted and observed BIS



(a) Indications of propofol sensitivity



(b) Indications of propofol tolerance

Figure 11: Although the RL agent was trained using a standardized patient prototype, the agent demonstrated good control in subjects deviating from the training model. In (a), the volunteer's observed BIS consistently measured below the predicted value, indicating the drug had greater effect than expected. In (b), the predicted BIS was consistently lower than measured, indicating the propofol dose yielded a higher BIS than anticipated.

values (Figure 11), the RL agent's ability to compensate for model mis-specification is highlighted. In the figure, Volunteer A demonstrated an apparent sensitivity to propofol. The observed hypnosis level consistently measured below the predicted value for most of the 30-minute period. Likewise, the RL agent compensated for an apparent propofol tolerance in Volunteer B. In this 30-minute period, the observed BIS consistently measured above the predicted value, indicating that the volunteer required more propofol than the population PK/PD models predicted. These observations suggest that the reinforcement learning process yielded a patient-specific control policy that may be applied to a general, variable population of volunteers with favorable results.

It should be noted that this RL implementation did not provide patient *adaptive* hypnosis, that is, no model parameters were adjusted during the control process. Rather, the agent exhaustively explored the discretized, bounded space of all possible $\text{BIS}_{\text{error}}$ and $\Delta\text{BIS}_{\text{error}}$ combinations during the exploring-start driven training. As a result, the agent formulated a control plan for all observable patient states, obviating a need for "on-the-fly" changes in its control policy. (Note that this style of exploration is limited to relatively small, discrete state spaces.) Adaptive closed-loop controllers in which model parameters are adjusted online have been studied in similar clinical control tasks (Ching et al., 2013; Shanechi et al., 2013; De Smet et al., 2008). Indeed, online reinforcement learning, a fixture in the intelligent systems literature, was a viable candidate for this application. However, a fixed-policy solution was preferred when seeking IRB approval for human study; likewise, the regulatory demands for any subsequent commercialization activities are lower when compared to an adaptive system. A convincing case is more easily made for a "safe and efficacious" system when the agent's control policy does not vary.

5.3 Limitations

The principal limitation of this study lies in its controlled nature. The human volunteers were healthy and mirrored those populations from which the PK/PD models were derived. Although the agent was challenged with credible intra-subject and inter-subject variation, it did not experience the full rigor of the intraoperative environment. In several instances of volunteer hypnosis, this limitation was realized with episodes of unanticipated natural sleep.

Because the bispectral index is an indirect measure of cortical activity, BIS is known to be affected by natural sleep (Nieuwenhuijs et al., 2002; Sleigh et al., 1999), as well as other conditions (including head trauma and hypothermia). In our study, conditions indicative of unanticipated natural sleep were observed after $\Delta\text{BIS}_{\text{target}}$ in some volunteers first receiving anesthesia at $\text{BIS}_{\text{target}}$=40. Figure 12 illustrates one instance in which the clinician directed a target change from 40 to 60 at $t \approx 64$ min. Since the desired target was higher than the subject's observed BIS, the agent correctly halted propofol infusion and waited for the volunteer to "recover" and awaken. Over the following ten minutes, the volunteer's predicted BIS rose as expected, but the volunteer's $\text{BIS}_{\text{measured}}$ remained near 40. $\text{BIS}_{\text{measured}}$ and $\text{BIS}_{\text{predicted}}$ increasingly diverged and the volunteer ultimately presented a predicted BIS near waking levels.

Given the ten-minute absence of propofol infusion and paradoxically low BIS measurements, the clinical team suspected the volunteer had transitioned from propofol-induced hypnosis to natural sleep. To continue the study and re-establish agent control, the clinician intervened with voice commands ("wake up", etc.) and a brief shoulder shake at $t \approx 74$ min. The volunteer's subsequent arousal was marked by a rapid convergence of $\text{BIS}_{\text{measured}}$ and $\text{BIS}_{\text{predicted}}$. As the volunteer awakened, the agent responded with propofol to reassert control at the new target of 60.

In summary, the volunteer shown in Figure 12 fell asleep shortly after the propofol infusion was interrupted—instead of waking as expected. In retrospect, this behavior was reasonable since our volunteer study lacked the usual surgical stimuli that would normally prevent natural sleep in the OR. Our volunteers were not subjected to persistent noxious

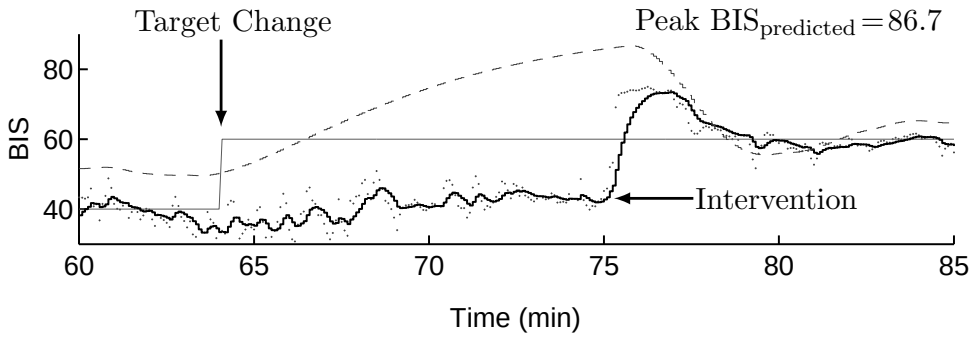


Figure 12: Our volunteer study modeled intraoperative hypnosis in a limited sense. Here, the agent stopped propofol infusion at the upward target change ($t \approx 64$ min). In response, the predicted BIS rose as the estimated effect-site concentration of propofol fell; however, $\text{BIS}_{\text{measured}}$ showed no corresponding increase. After ten minutes of no infusion, the subject failed to achieve a BIS > 40, although the predicted BIS approached waking levels. To continue study, the clinician intervened with rousing events (shoulder shake, voice commands, etc.). The volunteer responded immediately, and control resumed. This behavior was seen in several volunteers experiencing a low-to-high target change and was attributed to an unplanned transition to natural sleep after the target change.

stimulus, nor did they experience the usual bustling, noisy conditions of the operating suite. The ease in which the anesthetist roused the supposedly sedated volunteer, as well as the rate in which the measured BIS converged to the predicted, support the premise of natural sleep.

During the course of this study, the clinical team observed presumed natural sleep in 5 of 8 volunteers experiencing low-to-high target transition. (No sleep-like behavior was observed in volunteers undergoing high-to-low transitions.) Closer inspection of the low-to-high transition data (Table 11) reveals an apparent bimodal distribution in the corresponding time metrics. These observations appear to be naturally clustered in two well-differentiated groups: a *fast* transitioning group (non-sleepers) and a *slow* transitioning group (sleepers) that required an additional 3.5 min to emerge from Target 40 (Table 12). Exploratory parametric and non-parametric statistical tests suggest that two distinct groups do exist, but the small sample counts do not permit strong inferencing. The argument for sleep classification was bolstered when we confirmed *post-hoc* that all volunteers in the *slow* group required clinician intervention in order to complete the low-to-high transition. No *fast* group volunteers showed similar need. These findings suggest that presumed natural sleep occurred frequently in upward transitioning volunteers, thereby revealing a limitation of this study. Accordingly, our favorable results should be extrapolated to surgical patients in an appropriately limited fashion.

Fast Cluster				
	Vol	T_{sp} (min)	T_{peak} (min)	T_{ss} (min)
	2	7.13	8.13	8.13
	13	6.58	7.58	7.58
	14	8.59	8.92	9.59
Mean ± SD		7.43 ± 1.04	8.21 ± 0.67	8.43 ± 1.04
Median (IQR)		7.13 (1.01)	8.13 (0.67)	8.13 (1.01)
Slow Cluster				
	Vol	T_{sp} (min)	T_{peak} (min)	T_{ss} (min)
	4	11.25	12.09	14.25
	6	11.19	11.94	16.61
	8	11.39	12.36	14.86
	10	10.79	12.54	14.2
	12	10.84	12.17	14.26
Mean ± SD		11.09 ± 0.26	12.22 ± 0.23	14.84 ± 1.03
Median (IQR)		11.19 (0.41)	12.17 (0.27)	14.26 (0.61)

Table 12: Cluster Analysis of Target 40 to 60 Timed Metrics

5.4 Future Directions

Given the favorable performance in both simulation and healthy human volunteers, it seems reasonable to evaluate the agent in a study of actual surgical patients to more completely assess the clinical utility of RL control. Evaluation under the full rigor of the intraoperative environment, along with varying conditions of patient health, should provide further insight into RL's applicability. It is important to note that the studied RL agent does not directly represent a closed-loop drug delivery system suitable for general clinical use. For example, the current iteration of the agent is not equipped to reliably manage a prolonged open-loop condition due to BIS input failure. As such, it is more appropriate to consider the agent to be one player in a greater, more robust system.

The agent's aptitude for managing propofol response deviating from the training model (i.e., unexpected volunteer tolerance or sensitivity to propofol) is also cause for additional study. Like other PK/PD models, the Schnider PK model and the Doufas PD models characterize propofol response in a narrow, idealized population. Some poorly modeled populations, such as the critically-ill or morbidly obese, gain the most from patient-specific drug administration.

Finally, it should be noted that the application of reinforcement learning to medicine is not limited to depth-of-anesthesia management. Other potential applications exist, such

as neuromuscular blockade, mechanical ventilation, and management of cardiovascular parameters, including heart rate, blood pressure, and cardiac output.

5.5 Improvements

This study demonstrates the feasibility of RL hypnosis control, but it cannot yet be positioned as the optimal solution to the problem. Some areas of improvement are readily identifiable. For example, most of the metrics indicate that the agent controlled hypnosis more proficiently at $\mathrm{BIS}_{\mathrm{target}} = 60$. The authors theorize that the performance difference can be attributed to the apparent "mass" of a heavily-dosed patient. Patients at deeper levels of hypnosis accumulate higher concentrations of propofol, and the sigmoidal BIS response approaches saturation at levels near 4 μg/ml (Figure 3). Thus, an ever-increasing amount of propofol is required to meaningfully change the observed BIS in this region of the dose response curve. Likewise, the propofol-saturated patient responds to the zero infusion rate more slowly as peripheral tissue reservoirs continue to dump propofol into the patient's bloodstream well after the agent has discontinued infusion. These factors muddle the agent's interpretation of its actions, impairing its ability to regulate the patient's BIS level. In the following discussion, we suggest a few possible approaches to improve control.

We anticipate that the non-linearity illustrated in Figure 3 can be handled more effectively if the agent considers the current BIS measurement as an input. This additional percept provides a cue to handle the gross slope changes occurring in the ranges [0,1] μg, [1,4] μg, and [4,15] μg and should improve control at deeper levels of hypnosis.

An improvement may also be realized if the agent's goals can be modified to better reflect clinical practice. In general, control engineers are rightly concerned with achieving target setpoints with limited adverse behaviors, such as overshoot and ringing. Few anesthetists are control engineers, and few surgeons would appreciate the agent's observed $\tilde{1}$2.5-min induction of anesthesia. The agent's current reward strategy (Equation 1) discourages overshoot and promotes a "soft landing" on target; however, the clinician takes a different approach. A bolus of propofol is given, the patient is quickly rendered unconscious, and then the anesthetist manages any overshoot as needed. In other words, the goal of anesthetic induction differs fundamentally from the goal of anesthetic maintenance. Indeed, the goals oppose one another in time and control accuracy. A more effective solution might involve two independent, cooperative agents in which one agent is used for induction, and the other for maintenance.

The RL agent might also be implemented more effectively. In pilot studies, undesirable oscillation in the volunteer's BIS was occasionally observed. The authors theorized that the software filters used to attenuate BIS measurement noise compounded the 2.5-min lag in propofol effect, causing the agent to "chase" the target in an oscillatory fashion. To counter this noise without exacerbating lag, fuzzy state classifiers replaced aggressive smoothing so that the agent might better classify the volunteer's hypnotic state. The fuzzy classifiers reduced the significance of transient fluctuation in the $\mathrm{BIS}_{\mathrm{error}}$ and $\Delta\mathrm{BIS}_{\mathrm{error}}$ signals, thereby improving control performance so that human study could proceed as planned. Note that the fuzzy classifiers were selected without a comprehensive survey of filtering techniques. Clinical trials are expensive, challenging affairs not amenable to interruption once begun.

Fortunately, state generalization methods, like fuzzy classifiers, are well-represented in the RL literature,[5] Sparse coding (Sutton, 1996) and neural networks (Tesauro, 1992) are recognized methods for state aggregation in RL. Perhaps more fittingly, state generalization can be "rolled into" the Q-learning algorithm; fuzzy Q-learning, like that reported by Bonarini et al. (2009), Glorennec and Jouffe (1997) and Berenji and Kehdkar (1992), as well as delayed Q-learning (Chapman and Kaelbling, 1991), appear to be logical next steps in the evolution of this research.

It should also be noted that the agent was implemented as a discounted, infinite-horizon task ($\gamma < 1$). As mentioned previously, the closed-loop hypnosis task lacked an explicit goal state since the agent was expected to minimize control error for an undetermined duration. Alternatives exist for reinforcement learning in such infinite-horizon problems. Techniques that maximize returns over a window of time, like R-learning (Mahadevan, 1996), may be viable candidates for improving control performance.

Finally, when RL has been applied to real-world control tasks, the problem is usually modeled as an Markov Decision Process (MDP). This approach assumes complete observability of system states and influences that govern transitions among those states. In reality, full observability can be reasonably expected only in toy problems. Fortunately, hidden influences may be ignored without great consequence in many applications, leaving unadorned MDPs sufficient for the control task. However, closed-loop control of propofol hypnosis is a textbook example of a partially observable control process (Russel and Norvig, 2002). The task relies on an imperfect measurement (the bispectral index of the EEG) of a poorly-defined quantity (patient consciousness). Therefore, we believe that techniques used to solve Partially Observable Markov Decision Processes (POMDPs) (Kaelbling et al., 1998) are relevant in future studies.

6. Conclusion

The RL agent demonstrated clinically-suitable performance in the closed-loop control of propofol-induced hypnosis in healthy human volunteers. In doing so, the agent provided generalizing control that compensated for varying degrees of intra-subject and inter-subject variation in propofol effect, suggesting that RL control can improve propofol delivery in the general surgical population, as well as populations lacking good PK/PD models. Furthermore, RL's success in this clinical control task establishes precedence and positions the method as a viable candidate for solving other challenging clinical problems. Yet, as promising as these results appear, no strong conclusions regarding RL's place in closed-loop anesthesia can be made until similar results are observed under actual intraoperative conditions.

Acknowledgments

The clinical portion of this study was funded by the Department of Anesthesiology, Perioperative and Pain Management, Stanford University School of Medicine; the technical

5. More traditional predictive filtering techniques, like the Kalman filter, remain viable candidates for state generalization but are not discussed here.

aspects were funded by the authors. The authors would like to thank the Stanford University School of Medicine operating room staff for their support, as well as Aspect Medical (now Covidien) for providing an A-2000 BIS monitor.

References

A R Absalom and G N C Kenny. Closed-loop control of propofol anaesthesia using Bispectral Index^{TM}: Performance assessment in patients receiving computer-controlled propofol and manually controlled remifentanil infusions for minor surgery. *Brit J Anaesth*, 90(6):737–41, 2003.

A R Absalom, N Sutcliffe, and G N C Kenny. Closed-loop control of anesthesia using Bispectral Index^{TM}: Performance assessment in patients undergoing major orthopedic surgery under combined general and regional anesthesia. *Anesthesiology*, 96(1):67–73, Jan 2002.

M E Ausems, C C Hug, Jr, D R Stanski, and A G Burm. Plasma concentrations of alfentanil required to supplement nitrous oxide anesthesia for general surgery. *Anesthesiology*, 65 (4):362–73, Oct 1986.

M S Avidan, L Zhang, B A Burnside, K J Finkel, A C Searleman, J A Selvidge, L Saager, M S Turner, S Rao, M Bottros, C Hantler, E Jacobsohn, and A S Evers. Anesthesia awareness and the bispectral index. *New Eng J Med*, 11(358):1097–1108, Mar 2008.

J M Bailey, C T Mora, and S L Shafer. Pharmacokinetics of propofol in adult patients undergoing coronory revascularization. *Anesthesiology*, 84:1288–97, 1996.

L Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proc. 12th International Conference on Machine Learning*, pages 30–37. Morgan Kaufmann, 1995.

L Barvais, I Rausin, J B Glen, S C Hunter, D D'Hulster, F Cantraine, and A d'Hollander. Administration of propofol by target-controlled infusion in patients undergoing coronary artery surgery. *J Cardiothorac Vasc Anesth*, 10(7):877–83, Dec 1996.

H R Berenji and P Kehdkar. Learning and tuning fuzzy logic controllers through reinforcements. *IEEE Transactions on Neural Networks*, 3(5):724–740, 1992.

M J Bloom, A Bekker, C V Seshagiri, and S D Greenwald. Changes in BIS variability reflect changes in remifentanil infusion during spinal surgery. Presented at the American Society of Anesthesiologists Annual Meeting, Oct 2008.

A Bonarini, A Lazaric, F Montrone, and M Restelli. Reinforcement distribution in fuzzy Q-learning. *Fuzzy Sets and Systems*, 160(10):1420–1443, 2009.

J Boyan and A W Moore. Generalization in reinforcement learning: Safely approximating the value function. In G Tesauro, D S Touretzky, and T K Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 369–376, Cambridge, MA, 1995. The MIT Press.

E Brown, R Lydic, and N Schiff. General anesthesia, sleep, and coma. *N Engl J Med*, 363 (27):2638–50, Dec 2010.

A Carregal, A Lorenzo, J A Taboada, and J L Barreiro. Intraoperative control of mean arterial pressure and heart rate with alfentanyl with fuzzy logic. *Rev Esp Anestesiol Reanim*, 47(3):108–113, Mar 2000.

D Chapman and L P Kaelbling. Input generalization in delayed reinforcement learning: An algorithm and performance comparisons. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1991.

S Ching, B M Westover, M Liberman, J J Chemali, J Kenny, K Solt, P L Purdon, and E N Brown. Real-time closed-loop control in a rodent model of medically induced coma using burst suppression. *Anesthesiology*, 119(4):848–860, Oct 2013.

A Dahaba. Different conditions that could result in the bispectral index indicating an incorrect hypnotic state. *Anesth Analg*, 101(3):765–73, Sep 2005.

P Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8:341–362, 1992.

T De Smet, M M R F Struys, M M Neckebroek, K Van den Hauwe, S Bonte, and E P Mortier. The accuracy and clinical feasibility of a new Bayesian-based closed-loop control system for propofol administration using the bispectral index as a controlled variable. *Anesth Analg*, 107:1200–1210, 2008.

A G Doufas, M Bakhshandeh, A R Bjorksten, S L Shafer, and D I Sessler. Induction speed is not a determinant of propofol pharmacodynamics. *Anesthesiology*, 101:1112–21, 2004.

D Ernst, G B Stan, J Goncalves, and L Wehenkel. Clinical data based optimal STI strategies for HIV; a reinforcement learning approach. In *Machine Learning Conference of Belgium and The Netherlands (Benelearn)*, pages 65–72, 2006.

D Ernst, M Glavic, F Capitanescu, and L Wehenkel. Reinforcement learning versus model predictive control: A comparison on a power system problem. *Trans Sys Man Cyber Part B*, 39(2):517–529, 2009. ISSN 1083-4419.

V Esmaeili, A Assareh, M B Shamsollahi, M H Moradi, and N M Arefian. Estimating the depth of anesthesia using fuzzy soft computation applied to EEG features. *Intell Data Anal*, 12(4):393–407, 2008.

S P Fitzgibbon, D M Powers, K J Pope, and C R Clark. Removal of EEG noise and artifact using blind source separation. *J Clin Neurophysiol*, 24(3):232–43, Jun 2007.

A E Gaweda, M K Muezzinoglu, A A Jacobs, G R Aronoff, and M E Brier. Model predictive control with reinforcement learning for drug delivery in renal anemia management. *Conf Proc IEEE Eng Med Biol Soc*, 1:5177–80, 2006.

A Gentilini, C Frei, A H Glattfelder, M Morari, and T Schnider. Identification and targeting policies for computer controlled infusion pumps. *Crit Rev Biomed Eng*, 28(1&2):179–185, 2000.

E Gepts. Pharmacokinetic concepts for TCI anaesthesia. *Anaesthesia*, 53(Suppl 1):4–12, Apr 1998.

P S Glass, M Bloom, L Kearse, C Rosow, P Sebel, and P Manberg. Bispectral analysis measures sedation and memory effects of propofol, midazolam, isoflurane, and alfentanil in healthy volunteers. *Anesthesiology*, 86(4):836–847, Apr 1997.

P Y Glorennec and L Jouffe. Fuzzy Q-learning. In *Proceedings of Fuzz-IEEE'97, Sixth International Conference on Fuzzy Systems*, volume 3, pages 659–662, 1997.

S D Greenwald and C E Rosow. BIS and EMG variability increase before somatic responses during surgery. Presented at the American Society of Anesthesiologists Annual Meeting, Oct 2006.

A Guez, R D Vincent, M Avoli, and J Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *IAAI'08: Proceedings of the 20th national conference on Innovative Applications of Artificial Intelligence*, pages 1671–1678. AAAI Press, 2008. ISBN 978-1-57735-368-3.

V Gullapalli. Learning control under extreme uncertainty. In Stephen José Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 327–334. Morgan Kaufmann, San Mateo, CA, 1993. URL `citeseer.nj.nec.com/312133.html`.

J O Hahn, G A Dumont, and J M Ansermino. Closed-loop anesthetic drug concentration estimation using clinical-effect feedback. *IEEE Trans Biomed Eng*, 58(1):3–6, Jan 2011.

T M Hemmerling, S Charabati, C Zaouter, C Minardi, and P A Mathieu. A randomized controlled trial demonstrates that a novel closed-loop propofol system performs better hypnosis control than manual administration. *Can J Anaesth*, 57(8):725–735, Aug 2010.

C Hu, W S Lovejoy, and S L Shafer. Comparison of some control strategies for three-compartment PK/PD models. *Journal of Pharmacokinetics and Biopharmaceutics*, 22 (6):525–550, 1994.

L P Kaelbling, M L Littman, and A W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

L P Kaelbling, M L Littman, and A R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

L A Kearse Jr., P Manberg, N Chamoun, F deBros, and A Zaslavsky. Bispectral analysis of the electroencephalogram correlates with patient movement to skin incision during propofol/nitrous oxide anesthesia. *Anesthesiology*, 81(6):1365–70, Dec 1994.

G N C Kenny and H Mantzaridis. Closed-loop control of propofol anaesthesia. *Brit J Anaesth*, 83(2):223–8, Aug 1999.

K Leslie, A R Absalom, and G N C Kenny. Closed loop control of sedation for colonoscopy using the Bispectral Index. *Anaesthesia*, 57(7):690–709, Jul 2002.

M Lindholm, S Träff, F Granath, S D Greenwald, A Ekbom, C Lennmarken, and R H Sandin. Mortality within 2 years after surgery in relation to low intraoperative bispectral index values and preexisting malignant disease. *Anesth Analg*, 108(2):508–512, Feb 2009.

M Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning (ML-94)*, pages 157–163, New Brunswick, NJ, 1994. Morgan Kaufmann.

N Liu, T Chazot, A Genty, A Landais, A Restoux, K McGee, P A Laloë, B Trillat, L Barvais, and M Fischler. Titration of propofol for anesthetic induction and maintenance guided by the bispectral index: Closed-loop versus manual control: A prospective, randomized, multicenter study. *Anesthesiology*, 104(4):686–695, April 2006.

N Liu, M Le Guen, F Benabbes-Lambert, T Chazot, B Trillat, D I Sessler, and M Fischler. Feasibility of closed-loop titration of propofol and remifentanil guided by the spectral M-Entropy monitor. *Anesthesiology*, 116(2):286–295, Feb 2012.

N Liu, O Pruszkowski, J E Leroy, T Chazot, B Trillat, A Colchen, F Gonin, and M Fischler. Automatic administration of propofol and remifentanil guided by the bispectral index during rigid bronchoscopic procedures: A randomized trial. *Can J Anaesth*, 60(9):881–887, Sep 2013.

S Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22(1–3):159–195, 1996.

J D Martín-Guerrero, F Gomez, E Soria-Olivas, J Schmidhuber, M Climente-Martí, and N V Jiménez-Torres. A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert Syst Appl*, 36(6):9737–9742, Aug 2009.

D M Mathews, L Clark, J Johansen, E Matute, and C V Seshagiri. Increases in electroencephalogram and electromyogram variability are associated with an increased incidence of intraoperative somatic response. *Anesth Analg*, 114(4):759–70, Apr 2012.

B L Moore. Intelligent control of closed-loop sedation in simulated ICU patients. Master's thesis, Texas Tech University, 2003.

B L Moore, E D Sinzinger, T M Quasny, and L D Pyeatt. Intelligent control of closed-loop sedation in simulated ICU patients. In *FLAIRS 2004*. AAAI Press, 2004.

B L Moore, L D Pyeatt, and A G Doufas. Fuzzy control for closed-loop, patient-specific hypnosis in intraoperative patients: A simulation study. In *Conf Proc IEEE Eng Med Biol Soc*, volume 1, 2009.

B L Moore, A G Doufas, and L D Pyeatt. Reinforcement learning: A novel method for optimal control of propofol-induced hypnosis. *Anesth Analg*, 112(2):360–367, Feb 2011a.

B L Moore, T M Quasny, and A G Doufas. Reinforcement learning versus proportional-integral-derivative control of hypnosis in a simulated intraoperative patient. *Anesth Analg*, 112(2):350–359, Feb 2011b.

E P Mortier, M M R F Struys, T De Smet, Y D I Versichelen, and G Rolly. Closed-loop controlled administration of propofol using bispectral analysis. *Anaesthesia*, 53(8): 749–754, Aug 1998.

P Myles, K Leslie, J McNeil, A Forbes, and M Chan. Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial. *Lancet*, 363(9423):1757–1763, 2000.

D Nieuwenhuijs, E L Coleman, N J Douglas, G B Drummond, and A Dahan. Bispectral index values and spectral edge frequency at different stages of physiologic sleep. *Anesth Analg*, 94(1):125–129, Jan 2002.

D A O'Hara, D K Bogen, and A Noordergraaf. The use of computers for controlling the delivery of anesthesia. *Anesthesiology*, 77(3):563–81, Sep 1992.

K T Olkkola, H Schwilden, and C Apffelstaedt. Model-based adaptive closed-loop feedback control of atracurium-induced neuromuscular blockade. *Acta Anaesth Scand*, 35(5):420–3, Jul 1991.

S Omatu, M Khalid, and R Yusof. *Neuro-control and its Applications*, chapter 4, pages 152–160. Advances in Industrial Control. Springer, 1996.

S Pilge, R Zanner, G Schneider, J Blum, M Kreuzer, and E F Kochs. Analysis of cerebral state, bispectral, and narcotrend indices. *Anesthesiology*, 104(3):488–494, Mar 2006.

I J Rampil. A primer for EEG signal processing in anesthesia. *Anesthesiology*, 89(4): 980–1002, Oct 1997.

M Renna, T Wigmore, A Mofeez, and C Gillbe. Biasing effect of the electromyogram on BIS: A controlled study during high-dose fentanyl induction. *J Clin Monit Comput*, 17 (6):377–81, Aug 2002.

A E Rigby-Jones and J R Sneyd. Pharmacokinetics and pharmacodynamics: Is there anything new? *Anaesthesia*, 67(1):5–11, Jan 2012.

H Röpcke, M Knen-Bergmann, M Cuhls, T Bouillon, and A Hoeft. Propofol and remifentanil pharmacodynamic interaction during orthopedic surgical procedures as measured by effects on bispectral index. *J Clin Anesth*, 13(3):198–207, May 2001a.

H Röpcke, B Rehberg, M Koenen-Bergmann, T Bouillon, J Bruhn, and A Hoeft. Surgical stimulation shifts EEG concentration-response relationship of desflurane. *Anesthesiology*, 94(3):255–113, Mar 2001b.

S Russel and P Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2002.

F Sahba, H R Tizhoosh, and M M A Salama. Application of reinforcement learning for segmentation of transrectal ultrasound images. *BMC Med Imaging*, 8(8), 2008.

T Sakai, A Matsuki, P F White, and A H Giesecke. Use of an EEG-bispectral closed-loop delivery system for administering propofol. *Acta Anesth Scand*, 44:1007–1010, 2000.

R H Sandin, G Enlund, P Samuelsson, and C Lennmarken. Awareness during anaesthesia: A prospective case study. *Lancet*, 355(9205):707–711, 2000.

J Schaublin, M Derighetti, P Feigenwinter, S Petersen-Felix, and A M Zbinden. Fuzzy logic control of mechanical ventilation during anaesthesia. *Brit J Anaesth*, 77(5):636–41, Nov 1996.

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

T W Schnider, C F Minto, S L Shafer, P L Gambus, C Andresen, D B Goodale, and E J Youngs. The influence of age on propofol pharmacodynamics. *Anesthesiology*, 90(6): 1502–16, Jun 1999.

H Schwilden, J Schüttler, and H Stoeckel. Closed-loop feedback control of methohexital anesthesia by quantitative EEG analysis in humans. *Anesthesiology*, 67(3):341–7, Sep 1987.

P S Sebel, T A Bowdle, M M Ghoneim, I J Rampil, R E Padilla, T J Gan, and K B Domino. The incidence of awareness during anesthesia: A multicenter United States study. *Anesth Analg*, 99:833–839, 2004.

F S Servin. TCI compared with manually controlled infusion of propofol: A multicentre study. *Anaesthesia*, 53(Suppl 1):82–86, Apr 1998.

M M Shanechi, J J Chemali, M Liberman M, K Solt, and E N Brown. A brain-machine interface for control of medically-induced coma. *PLOS Compu Biol*, 9(10):1–17, Oct 2013.

J C Sigl and N G Chamoun. An introduction to bispectral analyis for the electroencephalogram. *J Clin Monitor*, 10(6):392–404, November 1994.

J W Sleigh, J Andrzejowski, A Steyn-Ross, and M Steyn-Ross. The bispectral index: A measure of depth of sleep? *Anesth Analg*, 88(3):659–661, Mar 1999.

M M R F Struys, T De Smet, S D Greenwald, A R Abasalom, S Bingé, and E P Mortier. Closed-loop controlled administration of propofol using bispectral analysis. *Anesthesiology*, 95(1):6–17, Jul 2001.

M M R F Struys, T De Smet, S D Greenwald, A R Absalom, S Bingé, and E P Mortier. Performance evaluation of two published closed-loop control systems using bispectral index monitoring: A simulation study. *Anesthesiology*, 100(3):640–7, Mar 2004.

M M R F Struys, M J Coppens, N De Neve, E P Mortier, A G Doufas, J F P Van Bocxlaer, and S L Shafer. Influence of administration rate on propofol plasma-effect site equilibration. *Anesthesiology*, 07(3):386–396, Sept 2007.

R Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In Touretzky, Mozer, and Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 1038–1044. The MIT Press, 1996.

R S Sutton and A G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

G Tesauro. Temporal difference learning of backgammon strategy. In *Proceedings of the International Conference on Machine Learning*, pages 451–457. Morgan Kaufmann, 1992.

D R Theil, T E Stanley, 3rd, W D White, D Goodman, P S Glass, S A Bai, J R Jacobs, and J G Reves. Midazolam and fentanyl continuous infusion anesthesia for cardiac surgery: A comparison of computer-assisted versus manual infusion systems. *J Cardiothorac Vasc Anesth*, 7(3):300–6, Jun 1993.

J N Tsitsiklas and B Van Roy. An analysis of temporal difference learning with function approximation. Technical Report LIDS-P-2322, Massachusetts Institute of Technology, 1996.

C Vanlersberghe and F Camu. *Modern Anesthetics (Handbook of Experimental Pharmacology)*, volume 182, chapter Propofol, pages 227–252. Springer, 2008.

J R Varvel, D L Donoho, and S L Shafer. Measuring the predictive performance of computer-controlled infusion pumps. *J Pharmacokinet Biopharm*, 20:63–94, Feb 1992.

C J C H Watkins. *Learning from Delayed Rewards*. PhD dissertation, Cambridge University, Computer Science Department, 1989.

W Wood. Variability of human drug response. *Anesthesiology*, 71(4):631–634, Nov 1989.

L Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

Y Zhao, M R Kosorok, and D Zeng. Reinforcement learning design for cancer clinical trials. *Stat Med*, 28(26):3294–315, Nov 2009.