A Reinforcement Learning Based Control Approach for Propofol-Induced Burst Suppression

Jason C Huang[1,2], Scott C Tadler[2,3], Brian J Mickey[3,2,1], Keith Jones[4,3], Kai Kuck[2,1]
Departments of [1]Biomedical Engineering, [2]Anesthesiology, [3]Psychiatry, [4]Neuroscience
University of Utah; Salt Lake City, UT

***Abstract* – High-dose propofol is being investigated for its potential antidepressant effect. Propofol is titrated to induce burst suppression, a specific EEG pattern. However, propofol is difficult to dose due to uncertainty in each patient's pharmacokinetics (PK) and pharmacodynamics (PD), and the lack of a commercially available monitor of propofol concentration. Clinicians currently infer the proper drug dose after observing the EEG response to the given dose. In this report we share our development of an automated controller to optimally administer propofol-induced burst suppression. We designed a deep deterministic policy gradient (DDPG) algorithm, which includes two deep neural networks and relates a 2-dimensional action space with a 3-dimensional state space. Our DDPG prototype did not satisfy our minimum training criteria. However, we share our diagnosis of current limitations in training a DDPG-based RL agent to administer propofol to PK-PD-simulated in silico patients. We also discuss potential solutions to improve RL agent training and performance.**

I. CLINICAL BACKGROUND

A recent open-label clinical trial at the University of Utah demonstrated potential efficacy in propofol's antidepressant effects [1], which are being further studied in a randomized controlled trial [2], along with revised dosing strategies for propofol. In the interventional group, high-dose propofol is administered to induce a specific burst suppression ratio (BSR), which is monitored and measured by the BIS™ Monitor (Medtronic, Dublin, Ireland), for a specific duration of time. Burst suppression is an EEG pattern with alternating periods of bursts and quiescence [3], which is similar to the EEG patterns observed in electroconvulsive therapy, and can alternatively be induced by anesthetics like propofol [4] or isoflurane [5] at higher doses.

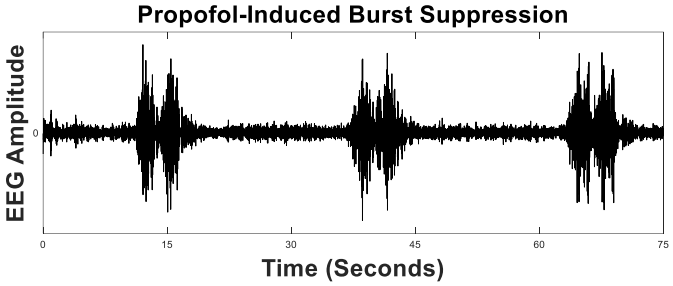

Figure 1. EEG recording of propofol-induced burst suppression during a high-dose treatment. The alternating periods of bursts and quiescence are segmented, then the ratio is determined by dividing the duration of suppressed EEG activity by the duration of the entire epoch of 60 seconds.

Titrating propofol to execute the treatment protocol is challenging, because patients' pharmacokinetics (PK) and pharmacodynamics (PD) vary [6, 7], and cannot be determined easily. Without technological assistance to administer propofol, clinicians are limited to their intuition and experience. There is neither a patient-specific nor standardized process to accurately and reliably control propofol-induced burst suppression (PIBS). This challenge impacts our clinical investigation of propofol's antidepressant effects.

II. TECHNICAL BACKGROUND

Dosing Based on PK-PD Modeling

PK-PD models can offer a way to conceptualize and estimate a patient's BSR response to administered propofol. Parameters from published PK models can be individualized to a specific patient by relating the propofol administered to the BSR observed in the patient's EEG. Effect site concentrations estimated based on the individualized PK model can then be used to estimate the patient's pharmacodynamics. Based on individualized PK-PD models, the propofol administration can be adjusted to achieve the desired levels and durations of burst suppression.

The main limitation of this approach is that individualizing model parameters would require a careful experimental design, which is not practical in the clinical

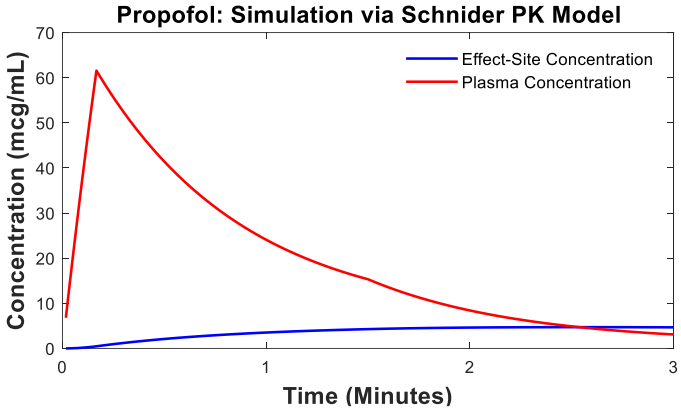

Figure 2. PK simulation of an administered propofol bolus, illustrating the accumulation of propofol in the central compartment (red) and in the effect-site compartment (blue). Drug accumulation in the effect-site lags behind the accumulation of drug concentration in plasma [7].

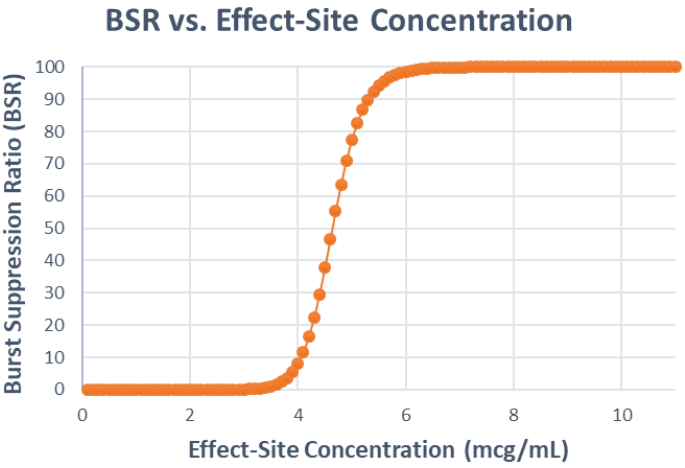

Figure 3. An example of a sigmoidal PD Hill Curve, relating effect-site concentrations to BSR. The concentration-response relationship is nonlinear and less sensitive to effect-site concentration changes at the BSR extremes. PD model parameters vary between and within patients.

setting. Neither plasma nor effect-site concentrations can be verified, because monitors for real-time propofol concentration monitoring are not available. Without any previous knowledge of how a particular patient responds to a drug, clinicians have to rely on population-based assumptions.

Without prior individualized estimations of a patient's PK-PD parameters, we can still apply PK-PD *principles* to guide decision making. We can also apply known, population-based *distributions* of PK-PD parameters and use machine learning to develop a controller that is robust enough to overcome the challenges of variability, uncertainty, and nonlinearity in PIBS.

Reinforcement Learning

Reinforcement learning (RL) is an intuitive goal-oriented control technique, which has demonstrated proficiency in solving challenging robotic tasks [8] and recently in controlling propofol anesthesia [9]. Without an explicit control algorithm or an individualized model of the patient, an RL "agent" may be able to learn optimal behavior on how to dose propofol and control BSR. The RL agent learns through experience from a reward function and observations from the environment.

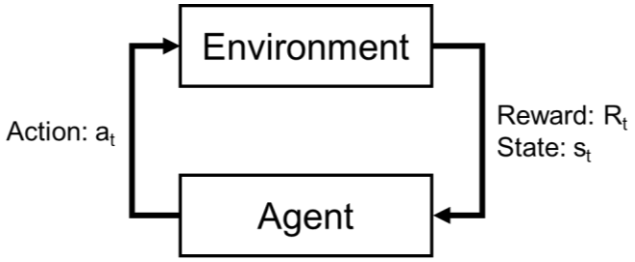

Figure 4. Block diagram illustrating the general structure of an RL agent's interaction with the environment. The Reward is the feedback that enables training and adjusts the determination of future actions.

Though the general structure of reinforcement learning is relatively simple, we must integrate the method with a simulated patient-environment; and properly structure the state space, action space, and reward function to effectively train the RL agent. We must also consider human factors, when deploying a RL agent in the real world.

For example, for commercializing an automated dosing system, it may be more practical if the clinician is kept in the control loop, due to regulatory concerns. In this case, clinicians would manually administer propofol, while the RL agent provides guidance to their decision making. The number of recommended dosing adjustments should be minimized and should not overburden the clinician.

III. OBJECTIVES

In this report, our objective is to develop and successfully train a RL agent on simulated patients of the same age, weight, and sex. Specifically, we seek to train the RL agent to optimally administer propofol and target a desired BSR. We hypothesize that our algorithm can

train a RL agent to reduce the average-absolute BSR error to <5%.

IV. METHODS

Create Simulated Patient

Published PK-model parameter distributions [7] and our group's own estimations of ke0 (mean ± SD of 0.136 ± 0.027 1/min), Hill coefficient (6.57 ± 1.70), and EC50 (7.40 ± 1.61 mcg/mL) were used to simulate the pharmacokinetics and -dynamics of 250 female patients with a height of 187 cm, age of 42 years, and weight of 96 kg.

Prior to any agent action, each simulated patient received a standard induction: bolus of 3 mg/kg and infusion of 300 mcg/kg/min.

Define RL State and Action Spaces

The RL state space is defined as:

1) BSR(t) – Target
2) |BSR(t) – Target| – |BSR(t-5 seconds) – Target|
3) Infusion(t)

The BSR Target was defined as 80% BSR.

The first state-variable tracks the proportional BSR error over each time step. The second state-variable tracks the change of the absolute BSR error over the last 5 seconds. The third state-variable tracks the represents infusion rate at time t. Though the infusion rate is not directly changed by the patient-environment, we believe that knowledge of the infusion rate can contribute to the RL agent's learning.

The RL action space is defined as:

1) Bolus, 0-100 mg
2) Infusion Rate, 0-400 mcg/kg/min

The time step was defined as 60 seconds.

We determined our action space to reflect the real-world decision making in our clinical investigation: In order to control BSR, clinicians either delivered a bolus dose of propofol or they adjusted the infusion rate. In our simulation, the RL agent applies both a bolus and infusion rate at each time step, within the ranges specified above.

Create a RL Agent

We applied a deep deterministic policy gradient (DDPG) algorithm [10] to create a reinforcement learning agent, which can handle continuous-multidimensional state and action spaces to solve complex problems

The DDPG-based RL agent is made up of two deep neural networks: the *actor* and the *critic*. When a DDPG-based RL agent processes its observations (states) from the environment (patient), the actor network determines a set of actions to apply to the environment (patient), while the critic network estimates the *Q-value* from the state-action combination. The Q-value is directly determined by the *reward function* and the discounted future rewards, according to the Bellman equation [11]. The actor is trained to maximize the long-term "reward," while the critic is trained to accurately estimate the Q-values from the combined state-action space.

The reward function is designed to steer the RL agent towards choosing actions based on specific states such that its propofol administration choices would achieve reaching the desired BSR target:

$$r(s_t, a_t) = \int_{\tau=t}^{\tau=t+step} | BSR(\tau) - Target | \, d\tau$$

This reward function is suitable in targeting a user-specified BSR during the induction and emergence phases of PIBS. Over the course of training, the function is designed to reduce the cumulative absolute BSR error.

Training the DDPG Agent

We followed the DDPG training algorithm described by *Lillicrap et al*. presented at ICLR 2016 [10], which specifies how the RL agent is trained, and how the actor and critic network weights are adjusted. The training was implemented in MATLAB (MathWorks, Natick, Massachusetts), using its Reinforcement Learning Toolbox (version 1.2).

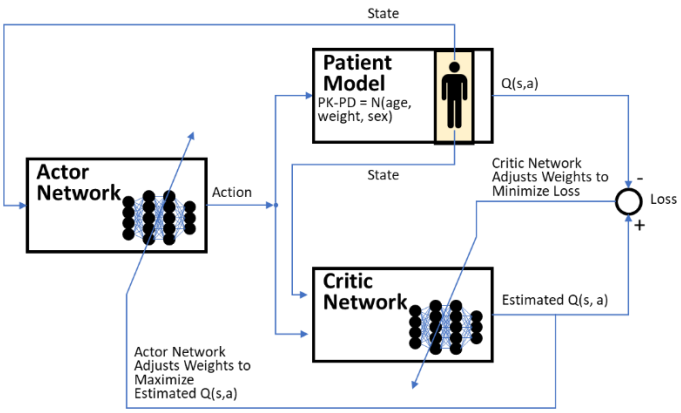

Figure 5. Block diagram illustrating the general structure of the DDPG algorithm, with the actor and critic networks interacting with the patient model. The critic network estimates the patient model's Q-value from the state-action combination. The Loss guides how the critic network adjusts its weights to optimize the critic's estimation. The actor network adjusts its weights according to 1) the gradient of critic output with respect to the applied action, and 2) the gradient of the actor output with respect to the actor weights, which together make up the overall gradient of the actor's performance based on Estimated Q(s,a) [11].

The RL agent was trained on each of the 250 simulated patients consecutively and for 120-steps per patient, where each step had a duration of 60-seconds. Ninety seconds after administering the standard induction dose, the RL agent began administering propofol and training its deep neural networks.

The *Ornstein-Uhlenbeck* noise process [12] was applied to the actor's action output before being applied to the patient and the critic network. Noise was used to promote exploration and avoid convergence toward local maxima. We selected a noise variance of 0.500 and variance decay rate of 10^{-5}, which reduces the noise variance after each time step throughout the entire training process.

After completing the training for each patient, we recorded the total reward over the 120-step training period. We also recorded a 5-patient moving reward average, which is based on the average of the total rewards of training five consecutive patients. Based on the hypothesis, the goal was for the average-absolute BSR error to be below 5% (criterion).

Because the reward function is a sum of absolute BSR error, we can calculate this average-absolute BSR error for each patient by dividing the total reward accumulated while training the system with that patient by the total time (120 time steps x 60 seconds per time step = 7200 seconds).

Our goal is to observe either a patient with an average-absolute BSR or an average from treating five consecutive patients (throughout any of the 250 patients in training) that satisfies the criterion of average-absolute BSR error below 5%

V. INITIAL RESULTS

After training on 250 patients, our DDPG agent did not meet the training criterion, nor demonstrate convergence toward the Critic's estimated discounted long-term reward. The best absolute-average BSR error was 7.65% BSR for a single patient, and 14.5% BSR for a 5-patient average. The DDPG agent also reported an average-absolute BSR error of over 20% BSR in 103 of the 250 patients (58.8%).

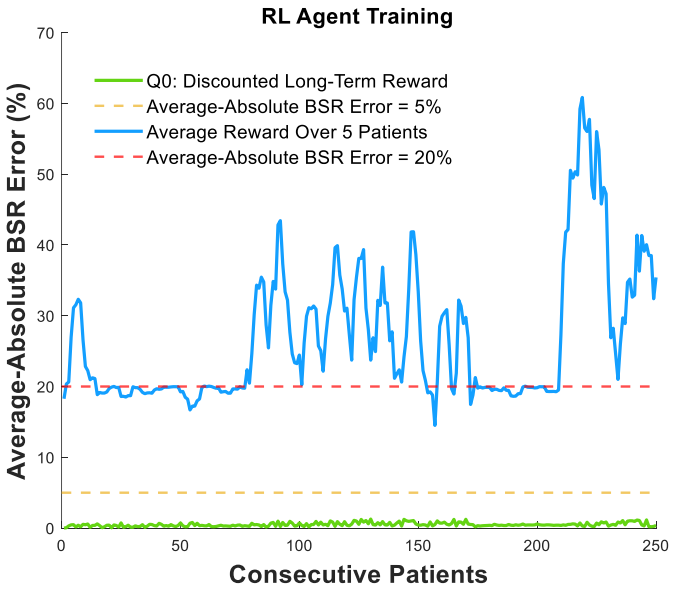

Figure 6 The 5-patient moving reward average (blue) illustrates the DDPG agent's performance as training progresses across patients. The agent's goal is to maximize the Q-value (calculated as the negative absolute BSR error and discounted future negative BSR errors) through each time-step and training for each patient rewards. Though noise and exploration in the action space can explain some fluctuations in performance, the DDPG does not demonstrate long-term improvement. An average-absolute BSR error of 20% is illustrated (red) to represent the possibility of the DDPG agent becoming "stuck" at the upper BSR extremes (~100% BSR) throughout the entire training for one patient. A successfully trained DDPG agent would achieve an average-absolute BSR error that decreases below the training criterion (yellow) and further converge toward the discounted long-term reward's average-absolute BSR error (green).

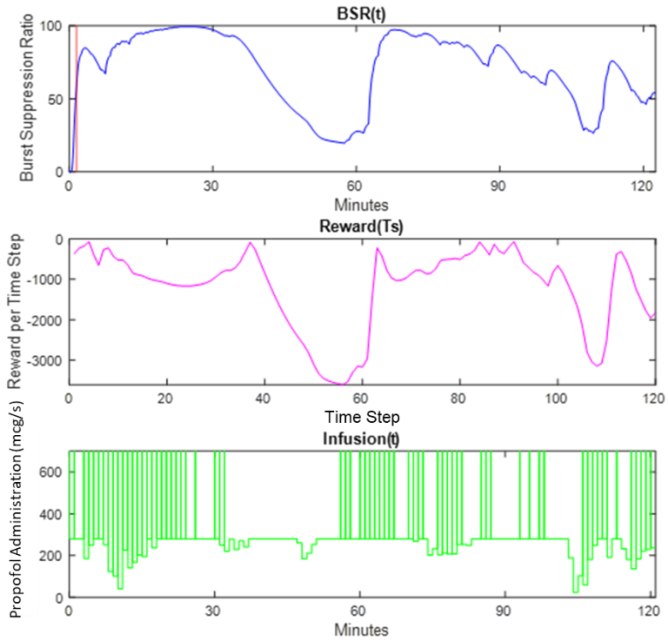

Figure 7 Example of training with patient #121, showing BSR, reward, and drug administration over time. The drug administration (green) includes a combination of a bolus and infusion, which affects the BSR (blue) and its corresponding reward (magenta). Even within a patient, we do not observe a consistent trend in improving the reward across time steps. The red vertical line on the BSR plot represents t = 90 seconds after the standard induction dose.

VI. DISCUSSION

For this initial design of DDPG agent, its patient-environment, and training structure, we reject our hypothesis, because the DDPG agent did not achieve a reward that surpasses the minimum training criteria, across the 250 patients it was trained on.

After observing the BSR signals and the rewards accumulated for each patient, we suspect that DDPG agent may be challenged in reducing BSR at the higher BSR extremes. Figure 7 blue reflects the inability to moderately decrease BSR (undershoot), while Figure 8 reflects the continual inability to significantly decrease BSR toward the target BSR of 80%. This can potentially be attributed to: 1) we cannot apply actions to directly and rapidly remove propofol from the patient-environment, and 2) the pharmacodynamic relationship between propofol concentration and BSR is nonlinear, while our reward function is linear.

In order to reduce BSR, we must reduce propofol delivery and rely on the patient to clear propofol through pharmacokinetics. This clearance is not as rapid as that of an administered bolus. In order to reinforce reductions in propofol delivery, longer time steps may be required to realize more significant changes in BSR and in the reward. The agent could benefit from the addition of memory and recurrent neural networks, as consecutive time steps of reduced propofol input may be required to effectively reduce BSR, as well as moderate the decrease in BSR over time.

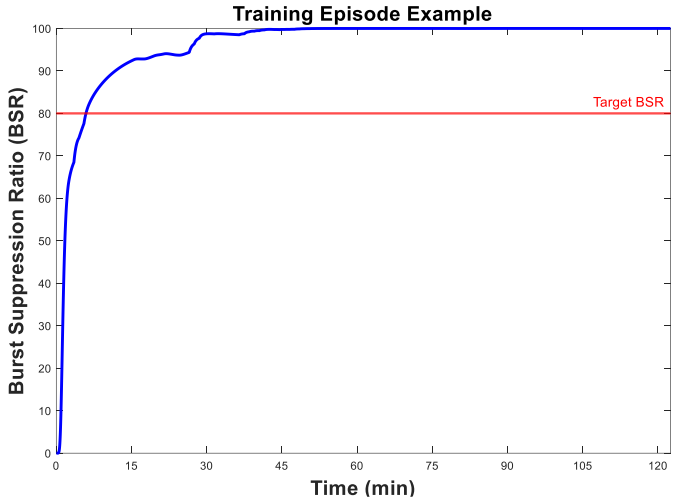

Figure 8. Example of a 120-step training with one patient, in which the RL agent is unable to decrease drug delivery, drug concentration, and BSR in the simulated training subject. The proposed reward function does not provide sufficient negative reinforcement to properly adjust the actor.

The reward function, as it was defined, might not have been properly "shaped" and can lead to a "vanishing gradient" problem in machine learning, in which the actor network is unable to adjust its weight, based on the feedback provided by the reward function. At the higher concentration and BSR extremes, the slope of the pharmacodynamic curve diminishes. Given that it is already difficult to reduce concentration, reducing BSR also becomes more difficult, while the current rewards function relies on changes in BSR magnitude. Thus, the changes in rewards across the action space would also diminish at the upper BSR extremes. This directly impacts the gradient of the policy's performance, which ultimately impacts how the actor network updates its weights [10].

VII. CONCLUSION

We created a PK-PD patient-model and integrated it into a reinforcement learning algorithm. Our current RL agent did not satisfy our minimum criterion

and was unable to converge toward higher rewards and a lower average-absolute BSR error.

Beyond modifying the reward function to improve the control of BSR accuracy precision, we must also train the RL agent to target a specific duration (12-15 minutes) of a specific BSR range (70-90%), as specified by the high-dose treatment protocol. We can also consider developing and deploying multiple agents with different goals, trained by different reward functions through the PIBS treatment. We must also consider how we integrate the RL agent with real-world clinical settings. We currently envision keeping the clinician in the control loop, and also seek to limit the number of dosing adjustments (e.g. no more than 5 adjustments) over each treatment, so that the clinician is not overburdened.

When we have demonstrated successful training in an RL agent, we plan to train the agents using patient-models with different sex, height, age, and weight combinations. Performance of a trained RL agent will be tested on a patient testing set that has not been seen during training. We also plan to apply a noise model to the BSR signal itself. We can evaluate the intra-patient, and inter-patient, and inter-treatment performance of a RL-based control approach for PIBS. If successful, these improvements in BSR control will directly support our clinical investigation of PIBS and other potential applications in anesthesia.

REFERENCES

[1] B.J. Mickey, S.C. Tadler, et al., "Propofol for Treatment-Resistant Depression: A Pilot Study," *International Journal of Neuropsychopharmacology*, vol. 21, no.12, pp.1079–1089, 2013.

[2] B. Mickey, "Neural and Antidepressant Effects of Propofol," *ClinicalTrials.gov*. [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT03684447.

[3] E. Niedermeyer E, "The burst-suppression electroencephalogram." *American Journal of Electroneurodiagnostic Technology*, vol. 49, no. 4, pp. 333–341, 2009.

[4] J. Bruhn, T.W. Bouillon, S.L. Shafer, "Onset of propofol-induced burst suppression may be correctly detected as deepening of anaesthesia by approximate entropy but not by bispectral index." *British Journal of Anaesthesia*, vol. 87, no. 3, pp. 505–507, 2001.

[5] S.H. Lisanby, "Electroconvulsive therapy for depression." *New England Journal of Medicine*, vol. 357, no. 19, pp. 1939–45, 2007.

[6] T.W. Schnider, et al. "The Influence of Age on Propofol Pharmacodynamics." *Anesthesiology*, vol. 90, no. 6, pp. 1502–1516, 1999.

[7] T.W. Schnider, et al., "The Influence of Method of Administration and Covariates on the Pharmacokinetics of Propofol in Adult Volunteers." *Anesthesiology*, vol. 88, no.5, pp. 1170–1182, 1998.

[8] J. Kober, et al., "Reinforcement learning in robotics: A survey." *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1239–1274, 2013.

[9] B.L. Moore, et al., "Reinforcement Learning: A Novel Method for Optimal Control of Propofol-Induced Hypnosis." *Anesthesia & Analgesia*, vol. 112, no. 2, pp. 360–367, 2011.

[10] T. P. Lillicrap, et al., "Continuous control with deep reinforcement learning," *International Conference on Learning Representations 2016*, San Juan, Puerto Rico.

[11] M. Lapan. *Deep Reinforcement Learning*, Cambridge, MA: MIT Press, 2019, pp. 99-108, 252-253, 410-412.

[12] G.E. Uhlenbeck, L.S. Ornstein, "On the theory of the brownian motion." *Physical Review*, vol. 36, no. 5, pp. 823–841, 1930.