Optimizing Drug Therapy with Reinforcement Learning: The Case of Anemia Management

Jordan M. Malof and Adam E. Gaweda

Abstract—**Optimal management of anemia due to End-Stage Renal Disease (ESRD) is a challenging task to physicians due to large inter-subject variability in response to Erythropoiesis Stimulating Agents (ESA). We demonstrate that an optimal dosing strategy for ESA can be derived using Reinforcement Learning (RL) techniques. In this study, we show some preliminary results of using a batch RL method, called *Fitted Q-Iteration*, to derive optimal ESA dosing strategies from retrospective treatment data. Presented results show that such dosing strategies are superior to a standard ESA protocol employed by our dialysis facilities.**

I. Introduction

ANEMIA is a universal consequence of End-Stage Renal Disease (ESRD). Untreated, anemia leads to a number of complications including cardiovascular disease, decreased quality of life, and increased mortality. Recombinant human erythropoietin (Epo) has been the mainstay treatment for anemia of ESRD for almost twenty years. The severity of anemia is clinically quantified by hemoglobin (Hgb) concentration in the blood. To standardize anemia treatment, regulatory authorities recommend maintaining Hgb between 10 and 12 g/dL in ESRD patients. To meet these guidelines, dialysis facilities attempt to determine the correct dose of Epo using Anemia Management Protocols (AMP). These AMPs are often merely verbal dose adjustment rules for different Hgb levels. As shown by recent evidence [1], Epo dosing patterns based on such protocols frequently lead to large variation in Hgb levels, both across the patient population as well as for within individual patients over time. The latter phenomenon has been labeled in the medical literature as "hemoglobin cycling" and associated with adverse health outcomes. From the clinical standpoint, finding an optimal dosing strategy for Epo poses a significant challenge since any such strategy must account for large and diverse patient populations.

As has been demonstrated previously [2], the challenge of drug dosing can be addressed through the use of automatic control. We have successfully demonstrated the application of control theory to anemia management using the principles of Model Predictive Control [3]. We and others have also investigated addressing this challenge using techniques rooted in the theory of Computational Intelligence and Machine Learning [4, 5]. In this work we explore Reinforcement Learning (RL) methods for the drug dosing problem and show that they can be used to derive Epo dosing strategies that match or outperform a standard AMP [6].

Reinforcement Learning is a collection of computational methods that mimic the trial-and-error-based knowledge acquisition used by humans and animals [7]. These methods represent the learning process as a set of interactions between an *agent* and its *environment*. The *actions* performed by the agent can affect the *state* of the environment, not only immediately but also over a long time. The adequacy of an agent's actions is judged by *rewards,* or *reinforcements*. The goal of the agent is to choose actions that will maximize its cumulative reward over some time interval.

In [4], we have demonstrated "in silico" that a RL method, called Q-learning, achieves a comparable performance to that of a standard AMP when applied to Epo dosing in an on-line mode. Other researchers have demonstrated that Q-learning can also be applied in an off-line mode to learn a dosing strategy from available treatment data [5, 8]. In this paper, we demonstrate the application of Q-learning to anemia management using an off-line approach. Specifically, we demonstrate how a relatively new RL method, called fitted Q-iteration [9], can be applied to derive an AMP from historical treatment data. We claim that this data-driven, RL-based AMP is superior to a standard clinical AMP developed by human experts.

The paper is organized as follows. We begin by explaining the most important concepts of RL theory, with special attention to fitted Q-iteration. Subsequently, we lay out our experimental design, including a description of the clinical data used and the relevant implementation details. We then present the experimental results and compare our approach to the current clinical standard. We conclude the paper with a short summary.

II. Methods

A. Reinforcement Learning

A Reinforcement Learning problem can be formulated as a Markov Decision Process (MDP). An MDP is a four-tuple: (S,A,P,R), where S is a set of states, A is a set of actions, P is a set of state transition probabilities, and R is a set of expected immediate rewards for reaching each state. The RL problem is to find a *policy*, a function $\pi(s)$ that specifies actions that the agent should choose when in each state s. The goal is to find a policy π that maximizes the expected sum of all future rewards, given by

J. M. Malof is with the Department of Electrical Engineering, Duke University, Durham, NC 27701, USA (e-mail: jordan.malof@duke.edu).

A. E. Gaweda is with the Department of Medicine, University of Louisville, Louisville, KY 40202, USA (phone: 502-852-0766, e-mail: adam.gaweda@louisville.edu).

978-1-4244-9637-2/11/$26.00 ©2011 IEEE

$$V = \sum_{k=0}^{\infty} \gamma^k r_{k+1}. \quad (1)$$

Here γ is the discount rate ($0 \leq \gamma < 1$) and determines the balance between the impact of the immediate and future rewards on the cumulative reward in (1). The agent learns the policy from past experiences (state-action pairs) by computing state-action values:

$$Q(s,a) = \sum_{s'} P_a(s,s')(R_a(s,s') + \gamma V(s')) \quad (2)$$

Here $P_a(s,s')$ is the probability of transition from state s to s', conditioned on action a. $R_a(s,s')$ is the reward associated with a transition from state s to s' and $V(s')$ is the value of state s'. The corresponding optimal policy π^* is defined as:

$$\pi^*(s) = \arg\max_a (Q(s,a)) \quad (3)$$

General RL methods based on the state-action value function (2) can be classified as *on-* or *off-policy*. On-policy methods derive the optimal policy by testing an active policy through trial and error and progressively refining it. On the other hand, off-policy methods learn the optimal policy from actions performed as a result of other policies. Q-learning [10] is the most popular off-policy method. We have previously demonstrated the application of Q-learning to on-line Epo dosing strategy estimation. In this work, we apply a modification of standard Q-learning, called *Fitted Q-Iteration* [9] to derive an optimal Epo dosing strategy from retrospective treatment data.

B. Cost Formulation of Reinforcement Learning

In its original formulation, the purpose of RL was to maximize a cumulative reward. RL can also be posed equivalently with the goal of minimizing costs (or penalties) c_t instead of maximizing rewards [11]. Using a cost formulation, the goal of RL becomes the minimization of total accumulated cost, given by

$$C = \sum_{k=0}^{\infty} \gamma^k c_{k+1}. \quad (4)$$

Here γ plays the same role as it did in (1). Using this cost formulation, (2) and (3) can be rewritten in the following way:

$$Q(s,a) = \sum_{s'} P_a(s,s')(c_a(s,s') + \gamma C(s')) \quad (5)$$

And

$$\pi^*(s) = \arg\min_a (Q(s,a)) \quad (6)$$

Our problem was naturally formulated in terms of costs rather than rewards, and so we used the cost formulation described above in our experimental design.

C. Fitted Q-Iteration

To apply Fitted Q-Iteration in this framework, data must be processed so that each datum is a 4-tuple vector that includes state, action, and cost information. At each time step, the algorithm uses the experience encoded in the data vector (s_t, a_t, c_t, s_{t+1}) to successively estimate the following recurrence relation:

$$\hat{Q}_{t+1}(s_t, a_t) = c_t + \gamma \min_{a \in A} Q_t(s_{t+1}, a) \quad (7)$$

Figure 1 shows an algorithmic representation of Fitted Q-Iteration.

Input: A set, $\mathcal{M}$, of 4-tuples where $\mathcal{M}_i = \{s_t^i, a^i, c^i, s_{t+1}^i\}$
Output: Q-value function estimator, $\hat{Q}$
Randomly initialize $\hat{Q}_0$ (k=0)
Repeat
{
  Generate set of training labels, $\mathcal{L}$, $\mathcal{L}^i = \hat{Q}_k(s_t^i, a^i)$
  Generate Pattern set
    $\boldsymbol{P} = \{(input^i, target^i), i = 1, \ldots, \#\mathcal{M}\}$ where:
      $input^i = \{s_t^i, a^i, \mathcal{L}^i\}$
      $target^i = c(s_t^i, a^i) + \gamma \min_a \hat{Q}_k(s_{t+1}^i, a)$
  Train estimator
    $\hat{Q}(\boldsymbol{P}) \rightarrow \hat{Q}_{k+1}$
  k = k + 1
}
Until stopping criteria met

Fig. 1. Algorithmic representation of Fitted Q-Iteration.

D. Data

We performed a retrospective observational cohort study of 209 patients receiving in-center hemodialysis at the Division of Nephrology, Department of Medicine, University of Louisville (Louisville, KY). All patients received hemodialysis treatment three times per week. Epo (Epoetin Alfa) was administered to patients intravenously during each treatment. Recorded data included monthly Hgb levels, weekly Epo dose, monthly transferrin saturation (TSat), serum albumin (Alb), Kt/V, quarterly serum ferritin (Ferr), and intact parathyroid hormone (PTH) levels. TSat and Ferr are surrogate markers for iron status. Serum Alb provides information about the nutritional status. Ferritin and Alb together can be used to diagnose inflammatory states. Kt/V is a measure of dialysis adequacy. Finally, PTH is a marker of metabolic bone disease. All these factors may affect patient's response to Epo.

Data were abstracted from Electronic Health Records as approved by the University of Louisville Institutional Review Board. To eliminate outliers and erroneous data entries, we selected data records where the Epo dose was between 0 and 100,000 Units per week, TSat between 0 and 100 %, Ferr level between 0 and 2,000 ng/mL, serum Alb between 0 and 5 g/dL, Kt/V between 0 and 2.5, and PTH between 0 and 2,000 pg/mL.

III. Experimental Design

A. Problem Definition

In our experiments, the *state* was represented by a patients current Hgb level and its rate of change:

$$s_t = \{Hgb, \Delta Hgb\} \quad (8)$$

Here, Hgb refers to the measured Hgb level in g/dL and ΔHgb refers to the calculated Hgb rate of change in g/dL per month. Additional state variables could reasonably be used in the state representation (e.g. Epo dose, TSat, Ferritin, Alb), however this simple scheme provides a good baseline for more elaborate future work. The *action* was represented by the change in Epo dose:

$$a_t = \Delta Epo \quad (9)$$

The treatment goal was to: (a) minimize a patient's time outside the target Hgb range (10-12 g/dL), (b) without rapid Hgb changes. The treatment goal was defined by a *cost function* as follows:

$$c_t = c_{t,Hgb} + c_{t,\Delta Hgb} \quad (10)$$

$$c_{t,Hgb} = \begin{cases} 0 & 10.5 \le Hgb \le 11.5 \\ 2 & 10.0 \le Hgb < 10.5 \\ 2 & 11.5 < Hgb \le 12.0 \\ 8 & otherwise \end{cases} \quad (11)$$

$$c_{t,\Delta Hgb} = \begin{cases} 0 & -1 \le \Delta Hgb \le 1 \\ 1 & -2 \le \Delta Hgb < -1 \\ 1 & 1 < \Delta Hgb \le 2 \\ 2 & otherwise \end{cases} \quad (12)$$

The cost component defined by (11) promotes Hgb levels close to the median of the target range 10-12 g/dL, which is recommended by the United States federal government [12]. The cost component defined by (12) penalizes Hgb changes greater than 1 g/dL per month, which may be associated with increased cardiovascular morbidity and mortality [13].

As a Q-value function approximation, we used a Feed-Forward Neural Network [11] with a single hidden layer, hyperbolic tangent transfer functions, and a linear output transfer function. The networks were trained by the resilient back-propagation algorithm [14]. To find the optimal network configuration, we varied two parameters: discount rate γ, and the number of hidden neurons η:

$$\gamma = \{\, 0.3\,, 0.4\,, 0.5\,, 0.6\,, 0.7\,, 0.8\,, 0.9\,\}$$

$$\eta = \{\, 2\,, 4\,, 6\,, 8\,, 10\,\}$$

B. Data Processing

To use Fitted Q-Iteration, information was extracted from the patient data in the form: $\{ s_t, a_t, c_t, s_{t+1} \}$. In this set, s_t is a state measurement from an arbitrary patient and a_t is the dose change received by that patient. Then c_t is the calculated cost of that state-action pair. This cost was calculated based on the resulting state, s_{t+1} using the cost function (10). Creating training patterns in this form, the dataset yielded 6251 patterns.

For each combination of the parameters (η,γ), a training session was conducted using a 4-fold cross-validation scheme. For a given set of parameters and a given training fold (out of the four), 10 networks were trained. Out of these 10 networks, the network with the best performance on the corresponding testing fold was retained. Performance, in this case, was measured by the metric given in section C by equation (13). This training scheme yielded one network per training fold, and four total networks for the whole training session. A block diagram illustrating this training procedure is shown in Figure 2.

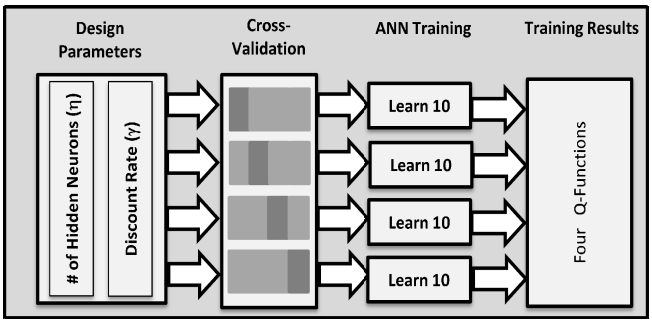

Fig. 2. Block diagram of experimental design

C. Performance Criteria

To test the performance of the dosing strategies derived from the trained networks, we adopted the approach proposed in [8]. For each i of the N data vectors in the test set, we calculated the following indicators:

1) $I_\pi(s_t^i, a_t^i)$ equals 1 if $a_t^i = \pi(s_t^i)$ and 0 otherwise, where π is the policy derived from the trained network. A policy recommendation $\pi(s_t^i)$ and action a_t^i were considered equivalent if they were within 500 units of each other.

2) $I_{HgbOut}(s_t^i)$ equals 1 if the Hgb level for state s_t^i was outside of the target range (10 – 12 g/dL) and 0 otherwise.

We then calculated the relative number of Hgb levels outside of the target range (10 – 12 g/dL) when using policy π, given by

$$\hat{S}_\pi = \frac{\sum_{i=0}^{N} I_\pi(s_t^i, a_t^i)\ I_{HgbOut}(s_t^i)}{\sum_{i=0}^{N} I_\pi(s_t^i, a_t^i)}. \quad (13)$$

We also calculated the estimated cumulative cost accrued when using policy π, given by

$$\hat{C}_\pi = \frac{\sum_{i=0}^{N} I_\pi(s_t^i, a_t^i)\ [c_t^i + \gamma \min_{a \in A} \hat{Q}(s_{t+1}^i, a)]}{\sum_{i=0}^{N} I_\pi(s_t^i, a_t^i)}. \quad (14)$$

The first performance measure estimates the clinical utility of the dosing strategy defined by policy π. The second performance measure estimates the utility of policy π in terms of the cost function (10). These performance measures were calculated for each policy using its testing data fold. In all cases, the testing dataset and training dataset were disjoint.

IV. RESULTS

The performance measures (13, 14) are reported in Tables I and II across all the testing parameters (γ - rows, η - columns). The table entries represent the average performance across the policies generated from each of the 4 validation folds. Table I shows that the largest percentage of Hgb on target is achieved with a discount rate of $\gamma = 0.6$. The results presented in Table II show that the lowest cost is achieved for a network with $\eta = 2$ hidden neurons. Based on these results, we suggest that the "best" overall combination of parameters is (2, 0.6). This "best" policy resulted in 83% of Hgb levels being within the target range, and an estimated cost of 1.2.

To benchmark these results, we evaluated a standard Epo dosing policy, based on the anemia management protocol used at our dialysis facility ("Standard Policy"). This policy was evaluated one time on the entire dataset (since no training was involved), again using measures (13, 14). The results from this test are reported in Table III along with the performance measures of the "best" RL Policy. The Standard Policy resulted in 54% of Hgb levels being within the target range. The "best" RL Policy resulted in almost a 50% relative improvement over the current standard in terms of this measure.

Figure 3 displays a surface plot of the Epo dose adjustments recommended by the best RL policy as a function of the states (Hgb, ΔHgb), and given a baseline (current) Epo dose of 12,000 Units per week. Figure 4 shows the same surface plot, but this time using the Epo dose adjustments recommended by the Standard Policy. The dose adjustments performed by the Standard Policy (Fig. 4) have a stepwise character, whereas the dose adjustments recommended by the RL Policy (Fig. 3) are gradual.

TABLE I
PERCENTAGE NEXT STATES OUTSIDE TARGET ($\hat{S}_\pi$)

Discount Rate (γ)	*Hidden Neurons* (η) 2	4	6	8	10
0.3	29	30	27	29	26
0.4	26	30	27	28	27
0.5	28	28	27	28	26
0.6	17	15	16	19	17
0.7	24	17	20	20	19
0.8	21	17	21	21	23
0.9	21	18	21	20	16

TABLE II
OVERALL COST FUNCTION ESTIMATE ($\hat{C}_\pi$)

Discount Rate (γ)	*Hidden Neurons* (η) 2	4	6	8	10
0.3	2.0	1.9	1.8	1.7	1.7
0.4	2.1	2.2	2.0	1.9	2.0
0.5	2.5	2.3	2.2	2.3	2.2
0.6	1.2	1.3	1.3	1.4	1.4
0.7	1.8	1.8	1.7	1.6	1.7
0.8	2.6	2.0	1.8	1.9	1.9
0.9	4.1	4.0	2.7	3.7	3.3

TABLE III
COMPARISON BETWEEN STANDARD AND RL POLICY

	Performance Measure $\hat{S}_\pi$	$\hat{C}_\pi$
Standard Policy	46	3.6
RL Policy	17	1.2

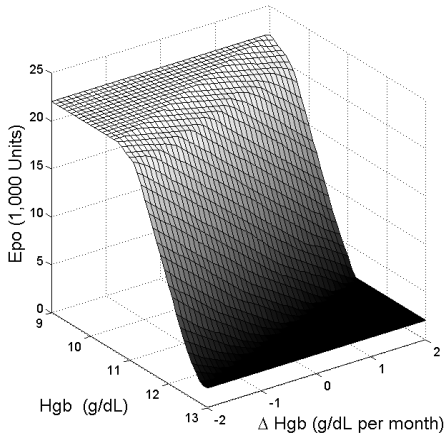

Fig. 3. Surface plot demonstrating new Epo dose (Epo) as a function of state variables (Hgb, ΔHgb) given a baseline dose of 12,000 Units, as recommended by the best policy estimated from treatment data using fitted Q-iteration ("RL policy").

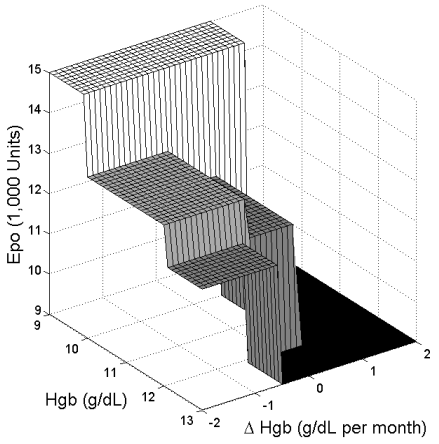

Fig. 4. Surface plot demonstrating new Epo dose (Epo) as a function of state variables (Hgb, ΔHgb) given a baseline dose of 12,000 Units, as recommended by the anemia management protocol ("Standard Policy").

V. Conclusions

This paper describes the application of a Reinforcement Learning method, called Fitted Q-Iteration, to the derivation of an optimal dosing strategy for Erythropoiesis Stimulating Agents in the management of anemia. Specifically, we have investigated the use of an Artificial Neural Network to approximate the Q-value function estimated from retrospectively collected treatment data. The experimental results indicate that this methodology has the potential to significantly improved clinical outcomes compared to the current clinical standard of care. Further research will focus on application of this method to concurrent dosing of multiple drugs and simultaneous optimization of multiple clinical outcomes in the anemia management of End Stage Renal Disease.

Acknowledgment

This work has been supported by National Institutes of Health Grant K25 DK072085.

References

[1] J. S. Berns, H. Elzein, R. I. Lynn *et al.*, "Hemoglobin variability in epoetin-treated hemodialysis patients," *Kidney Int,* vol. 64, no. 4, pp. 1514-21, Oct, 2003.

[2] R. Hovorka, V. Canonico, L. J. Chassin *et al.*, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiol Meas,* vol. 25, no. 4, pp. 905-20, Aug, 2004.

[3] A. E. Gaweda, A. Jacobs, G. R. Aronoff *et al.*, "Model predictive control of erythropoietin administration in the anemia of ESRD," *American Journal of Kidney Diseases,* vol. 51, no. 1, pp. 71-79, Jan, 2008.

[4] A. E. Gaweda, M. K. Muezzinoglu, G. R. Aronoff *et al.*, "Individualization of pharmacological anemia management using reinforcement learning," *Neural Netw,* vol. 18, no. 5-6, pp. 826-34, Jun-Jul, 2005.

[5] J. D. Martin-Guerrero, F. Gomez, E. Soria-Olivas *et al.*, "A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients," *Expert Systems with Applications,* vol. 36, no. 6, pp. 9737-9742, Aug, 2009.

[6] A. E. Gaweda, "Improving management of Anemia in End Stage Renal Disease using Reinforcement Learning." pp. 953-958.

[7] R. S. Sutton, and A. G. Barto, *Reinforcement learning : an introduction*, Cambridge, Mass.: MIT Press, 1998.

[8] J. Pineau, A. Guez, R. Vincent *et al.*, "Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach," *Int J Neural Syst,* vol. 19, no. 4, pp. 227-40, Aug, 2009.

[9] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-based batch mode reinforcement learning," *Journal of Machine Learning Research,* vol. 6, pp. 503-556, Apr, 2005.

[10] C. J. C. H. Watkins, and P. Dayan, "Q-Learning," *Machine Learning,* vol. 8, no. 3-4, pp. 279-292, May, 1992.

[11] M. Riedmiller, "Neural Fitted Q-iteration - First Experiences with a data efficient neural reinforcement learning method," in 16th European Conference on Machine Learning, 2005, pp. 317-328.

[12] "KDOQI Clinical Practice Guidelines and Clinical Practice Recommendations for Anemia in Chronic Kidney Disease," *Am J Kidney Dis,* vol. 47, no. 5 Suppl 3, pp. S11-145, May, 2006.

[13] A. Besarab, W. K. Bolton, A. R. Nissenson *et al.*, "The Normal Haematocrit Trial in dialysis patients with cardiac disease," *Nephrol Dial Transplant,* vol. 14, no. 8, pp. 2043-4, Aug, 1999.

[14] M. Riedmiller, and H. Braun, "A Direct Adaptive Method for Faster Backpropagation learning: the RPROP Algorithm," in IEEE Intl. Conference on Neural Networks, 1993, pp. 586-591.