Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Neural Networks 18 (2005) 826–834

Neural Networks

www.elsevier.com/locate/neunet

2005 Special Issue

Individualization of pharmacological anemia management using reinforcement learning☆

Adam E. Gaweda[a,*], Mehmet K. Muezzinoglu[b], George R. Aronoff[a], Alfred A. Jacobs[a], Jacek M. Zurada[b], Michael E. Brier[a,c]

[a]*Department of Medicine, University of Louisville, Louisville, KY 40292, USA*
[b]*Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292, USA*
[c]*Department of Veteran Affairs, Louisville, KY 40202, USA*

Abstract

Effective management of anemia due to renal failure poses many challenges to physicians. Individual response to treatment varies across patient populations and, due to the prolonged character of the therapy, changes over time. In this work, a Reinforcement Learning-based approach is proposed as an alternative method for individualization of drug administration in the treatment of renal anemia. *Q*-learning, an off-policy approximate dynamic programming method, is applied to determine the proper dosing strategy in real time. Simulations compare the proposed methodology with the currently used dosing protocol. Presented results illustrate the ability of the proposed method to achieve the therapeutic goal for individuals with different response characteristics and its potential to become an alternative to currently used techniques.

Keywords: Reinforcement learning; Drug dosing; Anemia management

1. Introduction

Drug administration in chronic conditions is a process of trial and error within a feedback loop. An initial drug dose is first selected as recommended by a standard reference. The patient is then observed for specific physiologic responses or adverse events. Subsequently, the clinician adjusts the dose following the observed state of the patient. If toxicity occurs, the dose amount is decreased. If an inadequate response is observed, the dose is increased. The trial and error process continues until a desired response is achieved.

Oftentimes, the relationship between the drug dose and the patient's response is complex. To facilitate drug administration, practitioners attempt to use protocols. Such protocols are developed from average responses to treatment in populations of patients. Nevertheless, achieving a desired therapeutic response on an individual basis is complicated by the differences within the population, as well as other concurrent medications and comorbidities, specific for each patient.

Reinforcement Learning (RL) is a methodology based on ideas from psychology that serves for control theory and stochastic optimization. It has a potential to become an effective tool for support of clinical decision making in patient care. (Bellman, 1983) described a general framework for applying Dynamic Programming (DP), a cornerstone methodology to RL, to pharmacotherapeutic planning using Pharmacokinetic and Pharmacodynamic (PK/PD) models. A pioneering demonstration of DP in pharmacotherapy can be found in (Buell, Jeliffe, Kalaba, & Sridhar, 1970). Other examples of using DP for pharmacotherapeutic planning include the works (Hu, Lovejoy, & Shafer, 1994a,b). (Schaeffer, Bailey, Shechter, & Roberts, 2004) reviewed various instances of medical application of Markov Decision Processes (MDP), the underlying control setting in RL. Most recently, (Moore, Sinzinger, Quasny, & Pyeatt, 2004) demonstrated how RL can be successfully employed in closed-loop control of patient sedation in an Intensive Care Unit.

Our previous work (Gaweda et al., 2005), which constitutes the origin of this paper, was aimed at discovering a complete administration policy for proper drug dosing

☆An abbreviated version of some portions of this article appeared in (Gaweda et al., 2005), published under the IEEE copyright.

* Corresponding author.

E-mail address: adam.gaweda@kdp.louisville.edu (A.E. Gaweda).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.06.020

during pharmacotherapeutic management of renal anemia. This was achieved by an on-policy RL method, SARSA, in which a patient model was probed by possibly non-optimal policies during an episodic learning process. Construction of such a policy requires sufficiently many occurrences of all possible state transitions, potentially causing over- or under-dosing. As a result, we showed that on-policy episodic RL tools can discover a useful dosing policy, as a product of a learning process, which may be however unacceptably long in real-time pharmacotherapy.

In this paper, we view the control problem at a lower level of generality, where the goal is to stabilize the Hemoglobin level within the target range of 11–12 mg/dl by evaluating reinforcements derived from state transitions. Due to the partially known, monotonic character of the dose-response relationship, we were able to reduce the Markov chain representation of the patient to a few representative states. In this way, the learning phase to reach an acceptable control can be shortened. To avoid probing the system by Suboptimal dosing policies during long training episodes, we utilize here a Q-learning mechanism (Watkins & Dayan, 1992) for evaluation of the state/action pairs. The proposed learning system determines the optimal drug dose using reinforcements, which are produced immediately after state transitions occurring within the patient dynamics during treatment. In contrast to our previous work where an RBF network was used for Q-table approximation, we use the RBF network here as an interpolator to identify the policy on the entire continuous state space.

The organization of the paper is as follows. Modeling patient dynamics and the drug dosing problem are presented in Section 2. Section 3 describes the use of a Markovian finite-state Q-learning method to achieve the control of the continuous-state patient dynamics. Experimental evaluation of the proposed approach is presented in Section 4. The results are also compared to those obtained using a simulated clinical protocol for anemia management. Concluding remarks and observations are discussed in Section 5.

2. Drug dosing problem

2.1. Patient dynamics

Anemia management is a typical control problem under uncertainty. The controlled quantity is the Hemoglobin level (HgB) and the control signal is the amount of Erythropoietin (EPO) administered by the physician. Iron stores in the body, determined by Transferrin Saturation (TSat), have an impact on the process of red blood cell production and are considered as an auxiliary state component. In this setting, a patient is viewed as a discrete-time dynamic system with the state space $\mathcal{H} \times \mathcal{S}$, where $\mathcal{H}$ and $\mathcal{S}$ are sets of valid HgB and TSat levels, respectively. We denote the control space, i.e. the set of valid EPO amounts, by $\mathcal{E}$. As the HgB and TSat measurements are performed monthly, the time index k denotes a month.

In the classical pharmacological framework, a patient's response is analyzed using a PK/PD compartment model containing a set of differential equations. In the case of the red blood cell production, called erythropoiesis, regular measurement of EPO concentration would be required to acquire all the information necessary to build a PK/PD model. Due to the expensive character of this procedure, alternative modeling methods, such as Artificial Neural Networks become a feasible option. In (Gaweda, Jacobs, Brier, & Zurada, 2003), a population-based neural network was proposed for dose–response modeling in anemia management. For the purpose of this study, we developed a 'subpopulation' approach. The underlying principle for this approach was the existence of several distinct response groups within a patient population. Each one of these groups was assumed to bear a unique dose–response relationship. Using fuzzy rules, a patient's response is first classified and subsequently a prediction of HgB level one-step ahead is performed using the following second-order model:

$$x_1[k+1] = \theta_1 u[k-1] + \theta_2 u[k] + \theta_3 u[k+1] + \theta_4 x_1[k-1] + \theta_5 x_1[k] + \theta_6 x_2[k] + \theta_0 \qquad (1)$$

where u is the control input (EPO), x_1 is the HgB, and x_2 is the TSat. The response is classified based on the six month average levels of HgB, TSat, and EPO. The proposed approach can be conveniently implemented using Takagi-Sugeno (TS) fuzzy model (Takagi & Sugeno, 1985).

Records of 186 patients at the Division of Nephrology, University of Louisville, were used to perform data-driven estimation of the TS model. The data were randomly divided into equally sized estimation (training) and evaluation (testing) sets, containing data of 93 patients each. For consistency, a total of 100 model estimations were performed using different patient selections for estimation and evaluation. Eventually, the following three-rule TS model was obtained:

R_1: If (*avg* EPO$_{6m}$, *target* HgB$_{6m}$, *norm* TSat$_{6m}$)
Then HgB$[k+1]=\Theta_1\zeta$
R_2: If (*avg* EPO$_{6m}$, *target* HgB$_{6m}$, *low* TSat$_{6m}$)
Then HgB$[k+1]=\Theta_2\zeta$
R_3: If (*high* EPO$_{6m}$, *low* HgB$_{6m}$, *low* TSat$_{6m}$)
Then HgB$[k+1]=\Theta_3\zeta$

In these rules, the subscript 6m denotes the six month average of the corresponding quantity, Θ_i are the parameter vectors of the predictive model (1), and ζ is the regressor vector:

$$\zeta = [\mathrm{EPO}[k-1], \mathrm{EPO}[k], \mathrm{EPO}[k+1], \mathrm{HgB}[k-1], \mathrm{HgB}[k], \mathrm{TSat}[k], 1]$$

Two rules (R_1, R_2) specify the HgB response for 'normal responders', i.e. patients who achieve 'target' HgB levels upon administration of 'average (avg)' EPO amount (ca. 12, 000 Units per week). These two rules cover 'normal responders' with 'low' and 'normal' TSat, respectively. The third rule (R_3) specifies the HgB response function for a group of patient, called 'poor responders'. These are patients who receive 'high' amounts of EPO yet their HgB level stays 'low'. The reason for using fuzzy sets to represent the response groups is due to the fact that patients in real life exhibit features typical for both groups to a certain degree. In other words, only very few patients can be classified strictly as a 'normal' or 'poor' responder.

In what follows, we assume x_2 to be a random process with normal distribution around mean $\bar{x}_2$ with a fixed variance σ^2_{TSat}:

$$x_2[k+1] \sim N(\bar{x}_2, \sigma^2_{\text{TSat}}), \tag{2}$$

and is bounded by 0 and 100. Since TSat is utilized in the prediction of HgB level, its random variation emulates the uncertainty in the process dynamics.

2.2. Problem statement

The control objective is to drive the HgB level (x_1) to and maintain within the target range 11–12 g/dl.

The only information utilized by the control method adopted in this work is the observed state transitions together with the control actions applied at each time step. In other words, the simulation-based control method described next does not require any model-specific information. The recursions (1) and (2) are used only for the purpose of simulating a patient's response.

3. Reinforcement learning approach to drug dosing

In this work, we cast the control problem of the preceding section as an RL problem, where gaining experience and improving the policy are considered as integrated subtasks to be achieved sequentially. The learning occurs in the form of immediate improvements in the drug dosing policy due to the experience gained by observing the HgB and EPO sequences of an individual patient.

As posed formally in the previous section, EPO dosing is a control problem with a discrete-time, continuous-state, second-order stochastic system. However, the learning system considered here operates on finite state and action spaces, so quantization of the observed state values to some representative levels is necessary. We denote the quantized HgB space as $\hat{\mathcal{H}}$ and assume that it consists of finite representative HgB levels equally spaced in the original HgB space $\mathcal{H}$. Similarly, $\hat{\mathcal{E}}$ denotes the set of quantization levels equally spaced in EPO space, and any action processed by the learning system is assumed to be already quantized to $\hat{\mathcal{E}}$.

Note that, in our formulation, TSat level is a pure normal random variable for each k and has no direct effect on the control objective. Then, x_1 is the sole state variable influencing the control objective and the transitions among the representative levels can be viewed as an MDP on $\hat{\mathcal{H}}$ governed by the actions in $\hat{\mathcal{E}}$ and the uncertainty due to x_2. In this setting, the learning system is expected to relate the representative HgB levels $\hat{\mathcal{H}}$ to some EPO doses in $\hat{\mathcal{E}}$ in an algebraic way, yielding a drug dosing policy (control law) valid only for $\hat{\mathcal{H}}$.

The learning system detailed below is responsible for calculating the optimal actions (EPO in 1000 Unit steps) to take at each representative HgB level. The amount of EPO to be applied is then calculated based on the actual (non-quantized) HgB level of the patient by an RBF network, which interpolates the drug dosing policy from the finite samples produced by the learning system.

The overall learning mechanism described in this section is illustrated in Fig. 1.

3.1. Q-learning system

A critical issue in the choice of a suitable RL tool is to decide whether to learn on or off-policy. Off-policy methods enable learning by observing the effects of a policy other than the one processed, so that probing the plant (the patient) with inadequate policies can be avoided. This makes off-policy methods appropriate tools for medical applications. To develop a drug dosing policy for an individual patient, we adopt here particularly the $Q(\lambda)$ algorithm, an off-policy RL process operating on finite-state/finite-action spaces.

The considered procedure evaluates all possible state/action pairs by maintaining an array $Q \in \mathbb{R}^{|\hat{\mathcal{H}}| \times |\hat{\mathcal{E}}|}$. Each entry of this Q table is interpreted as the unique measure of preferability of the associated (s, a) pair among all possible state/action pairs. Q table is indeed a simulation-based

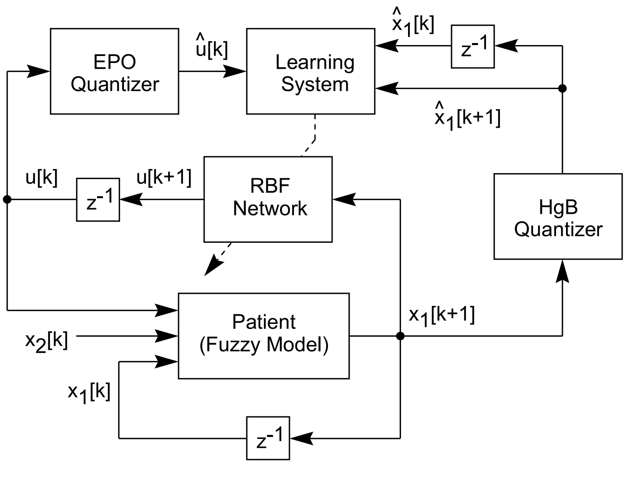

Fig. 1. Block diagram of the control process.

estimate of the optimal values, which satisfy the Bellman equation:

$$\hat{Q}(s,a) = E[g(s,s') + \gamma\hat{Q}(s',a')]$$

for all $s \in \hat{\mathcal{H}}$ and $a \in \hat{\mathcal{E}}$, where $g(\cdot,\ \cdot)$ is the immediate reward associated to the state transition $s \to s'$, $\gamma \in [0,1]$ is a fixed discount factor and the expectation is taken with respect to the distribution of all possible successors (s', a') of state s.

In this setting, the learning system (see Fig. 1) observes the current quantized state $\hat{x}_1[k]$, the quantized action (EPO dose) $\hat{u}[k]$, and the quantized successor state $\hat{x}_1[k+1]$. Given this information, a rewarding mechanism embedded in the learning system evaluates $g(\hat{x}_1[k], \hat{x}_1[k+1])$ based on the immediate contribution of the current state transition toward the control goal. This real valued reward is the sole information reflecting the control objective to the learning process and is used to update the entry $Q(\hat{x}_1[k], \hat{u}[k])$. In this study, we use the following reward function to assess the immediate benefit of the state transition $s \to s'$ to stabilize the state variable within the target range [11.0,12.0]:

$$g(s,s') = \begin{cases} 2, & \text{if } 11.0 \le s' \le 12.0 \\ -1, & \text{if } \{s < 11.0 \text{ and } s' < s\} \\ & \text{or } \{s > 12.0 \text{ and } s' > s\} \\ 1, & \text{otherwise} \end{cases}$$

Note that $g(\cdot,\ \cdot)$ described above assigns the highest reward to state transitions into the target range. Any transition that misses the target receives the lowest reward, i.e. penalty. Obviously, other reward functions consistent with the same control objective could be used (Gaweda et al., 2005).

The action $\hat{u}[k]$ is then related to the current quantized state $\hat{x}_1[k]$ through the algebraic policy $p(\cdot) : \hat{\mathcal{H}} \to \hat{\mathcal{E}}$, which is iterated as described below.

In particular, it can be shown that the temporal difference

$$\delta[k] = g(\hat{x}_1[k], \hat{x}_1[k+1]) + \gamma Q(\hat{x}_1[k+1], \hat{u}[k+1]) \\ - Q(\hat{x}_1[k], \hat{u}[k])$$

associated to the state transition $\hat{x}[k] \to \hat{x}[k+1]$ due to the action $u[k]$ is a correction on the estimate $Q(\hat{x}_1[k], \hat{u}[k])$ of the actual value state/action pair $\hat{Q}(\hat{x}[k], \hat{u}[k])$. For each transition $\hat{x}[k] \to \hat{x}[k+1]$ encountered due to $\hat{u}[k]$, the $Q(\lambda)$ algorithm performs the update

$$\hat{Q}(\hat{x}_1[k], \hat{u}[k]) \leftarrow \hat{Q}(\hat{x}_1[k], \hat{u}[k]) + \nu\delta[k](1 + e(\hat{x}_1[k], \hat{u}[k])), \tag{3}$$

where ν is a sufficiently small learning rate and $e(\mathbf{x}[k], \hat{u}[k]) \ge 0$ denotes the eligibility of the state/action pair $(\mathbf{x}[k], \hat{u}[k])$ in this correction.

After this correction, before proceeding with the next transition, eligibility of the current state/action pair is first updated as

$$e(\hat{x}_1[k], \hat{u}[k]) \leftarrow 1 + e(\hat{x}_1[k], \hat{u}[k])$$

and then the entire e array is iterated as

$$e \leftarrow \nu\lambda e,$$

$\lambda \in [0,1]$ is a parameter of the algorithm. Where λ is small the state/action pairs loses rapidly their eligibilities to update the Q entries. So the frequency of encountering a particular state/action pair in the trajectory becomes a less important effect in the update of the associated Q entry. For $\lambda = 1$, all encountered state/action pairs are treated equal in terms of their eligibility in the update of Q table, irrespective of the order of their occurrence in the trajectory.

After the Q and e updates for each state transition, the final step performed by the algorithm is the extraction of the policy based on the resulting Q table:

$$p(\hat{x}) = \underset{u \in \mathcal{E}}{\operatorname{argmax}}\, Q(\hat{x}, u).$$

This particular policy determined merely as the maximum element of the associated row of Q is called a greedy policy.

For diminishing learning constant ν and for $\lambda \in [0,1]$, the iteration on the policy performed after each state transition based on the last Q update converges to an optimal policy, provided that the visits to each state/action pair are sufficiently frequent (Tsitsiklis, 1994).

The last condition on the convergence to the optimal policy is satisfied by a modified version of the greedy policy, namely ε-greedy policy (Sutton & Barto, 1998):

$$p(\hat{x}) = \begin{cases} \max_{a \in \mathcal{E}} Q(\hat{x}, a), & z > \epsilon \\ \text{an arbitrary element of } \mathcal{E} & \text{otherwise} \end{cases}, \tag{4}$$

where z is a random variable distributed uniformly within [0,1], and $\epsilon \in [0,1]$ is a parameter of the learning algorithm. Note that the policy updated in this way contributes to the exploration in the search of the optimal policy in a different way than the dynamic uncertainty x_2 does.

3.2. RBF policy network

Approximating the dynamic programming table using artificial neural networks has been proven to be an effective way of handling large decision making problems (Bertsekas & Tsitsiklis, 1996). There are basically two points where the aid of neural networks to this kind of control process could be necessary. The first one is a compact parametric representation of the Q array, which turns out to be essential as the cardinality of the state space expands. However, in our particular setting, there are only a few representative states and actions, so maintaining explicitly the Q array here is computationally feasible. On the other hand, a connectionist approximation scheme could serve for a more critical purpose in this control process, i.e. to identify the control policy over the entire state space $\mathcal{H}$.

Since the resulting policy (4) obtained by the $Q(\lambda)$ algorithm is valid only for the quantized states $\hat{x} \in \hat{\mathcal{H}}$, to be applicable to our patient model, $p(\cdot)$ needs to be generalized to the cover $\mathcal{H}$ by means of an interpolator. We note that RBF networks have been proven to be effective tools for generalizing the control law (Sanner & Slotine, 1992 in similar cases).

We propose an RBF network with $|\hat{\mathcal{H}}|$ Gaussian RBF nodes centered at the representative states. The actual EPO dose to be applied for a given HgB level $x_1[k]$ is then determined by this policy network as

$$u[k] = [p(s_1)\cdots p(s_{|\hat{\mathcal{H}}|})]\begin{bmatrix} \exp\left(-\dfrac{x_1[k]-s_1}{\sigma}\right) \\ \vdots \\ \exp\left(-\dfrac{x_1[k]-s_{|\hat{\mathcal{H}}|}}{\sigma}\right) \end{bmatrix}$$

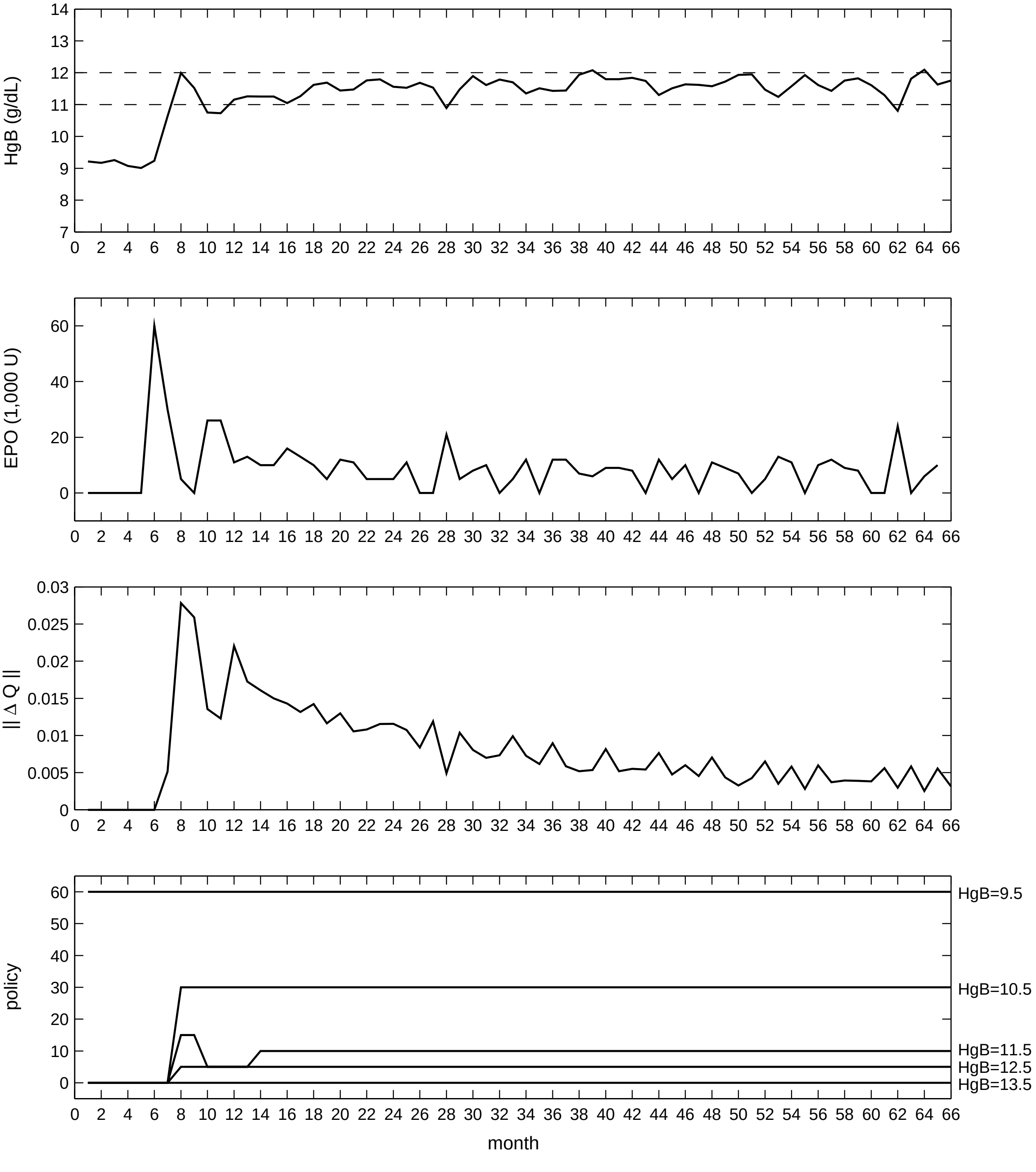

Fig. 2. HgB level (top), EPO dose (second from top), the magnitude of the Q-table updates (third from the top), and policy evolution (bottom) for an individual 'normal responder' as performed by Q-learning with RBF Policy Network.

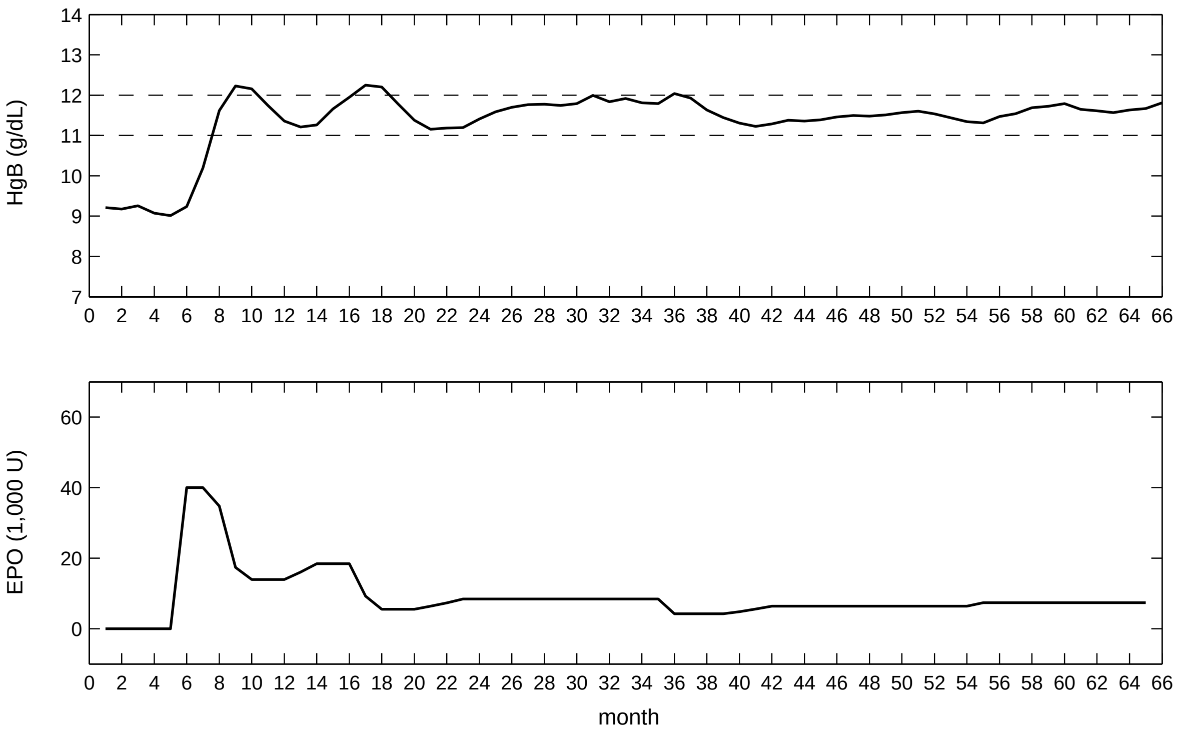

Fig. 3. HgB level (top) and EPO dose (bottom) for an individual 'normal responder' as performed by AMP.

where $\{s_i\}_{i=1}^{|\hat{\mathcal{H}}|}$, are the representative states, i.e. the elements of $\mathcal{H}$.

Since the representative states are equally spaced in $\mathcal{H}$, the spreads σ of the Gaussian nodes can be picked as

$$\sigma = \frac{d}{2},$$

so that the outputs of all nodes add up to approximately 1 for all points in $\mathcal{H}$, where d is the distance between two consecutive representative HgB levels in $\hat{\mathcal{H}}$. This enables assigning valid degrees of membership to the representative levels, so this mechanism gives an acceptable EPO dose in the form of the weighted sum of the actions imposed by the discrete policy. In this way, the RBF network plays a critical role by implementing the algebraic policy in the proposed drug dosing scheme.

4. Experimental results

To perform an experimental evaluation of the proposed method, we created an artificial group of 200 patients. Out of this group, the first 100 were typical for 'normal responders', while the remaining 100 were typical for 'poor responders.' For each individual patient, a trajectory of EPO, TSat, and HgB was generated over 6 months. To create these trajectories, we randomized data from actual individuals representative for each response group in our patient data base.

We simulated 5 years of anemia management for each patient group using the following two methods:

- Q-learning with RBF Policy Network
- Anemia Management Protocol (AMP)

Anemia Management Protocol is a numerical implementation of an EPO administration protocol which is currently used at the Division of Nephrology. This last simulation was performed to establish a 'gold standard' to which the results obtained by Q-learning can be compared. It must be pointed out that the AMP uses a mechanism for determination of EPO dose which is quite different and more involved than that used in our Q-learning based scheme. Furthermore, the administration strategy implemented by AMP is fixed a

Table 1
Simulation statistics for Q-learning

Response group	Normal	Poor
HgB level	11.59 (11.12, 12.04)	11.16 (10.76, 11.55)
HgB variability	0.29 (0.15, 0.42)	0.74 (0.52, 0.95)
Total EPO (1000 U)	589.29 (344.56, 834.02)	1145.25 (926.61, 1363.88)

Table 2
Simulation statistics for AMP

Response group	Normal	Poor
HgB level	11.66 (11.56, 11.78)	11.51 (11.35, 11.67)
HgB variability	0.32 (0.22, 0.41)	0.67 (0.49, 0.84)
Total EPO (1000 U)	610.57 (356.91, 864.23)	1075.39 (942.50, 1208.28)

priori, as opposed to the one used in Q-learning, which evolves in time. The dose selection procedure, as implemented in AMP, can be shortly described by the following expression:

$$\Delta\text{EPO}[k] = F[\text{HgB}[k-1], \text{HgB}[k-2], \text{HgB}[k-3], \text{EPO}[k-1]]$$

This is a higher order dynamic system, as opposed to a simple algebraic policy representation in (4). In the case of Q-learning, the state (HgB) was quantized into 5 equally sized intervals with medians at: 9.5, 10.5, 11.5, 12.5, 13.5 g/dL. These four values constituted the finite set $\hat{\mathcal{H}}$ explained in Section 3. The finite action set used by the learning system was defined as $\hat{\mathcal{E}} = \{0, 5, 10, \ldots, 60\}$. Due to physical constraints, we rounded the action values to an

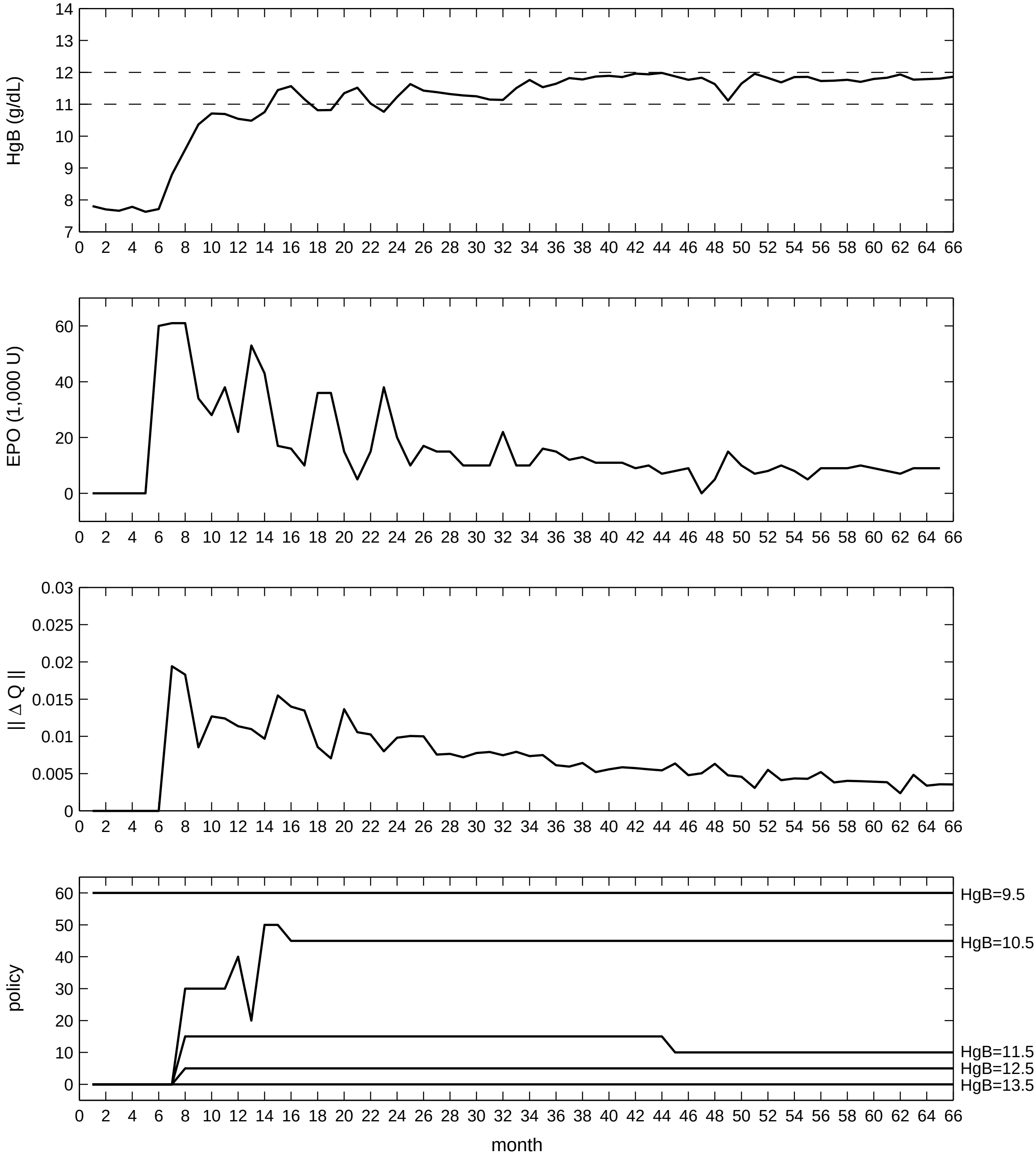

Fig. 4. HgB level (top), EPO dose (second from top), the magnitude of the Q-table updates (third from the top), and policy evolution (bottom) for an individual ‘poor responder’ as performed by Q-learning with RBF Policy Network.

integer (1,000 EPO Units is the lowest dose increment currently used).

In each simulation, the treatment was started after the sixth month. The Q-table was initialized using a 'best guess' method such that the most viable policy was used in the first step. The 'best guess' policy used in the simulations was as follows:

HgB (g/dl)	9.5	10.5	11.5	12.5	13.5
EPO (1000 U.)	60	30	15	5	0

When a new patient comes in, they cannot be classified immediately to a response group. This information is obtained as the treatment progresses. Consequently, using one sound and common initial policy and tailoring it for the individual patient during treatment is a viable solution. Thus, we used the same initial policy for both response groups. Furthermore, we inhibited updates to the policy entries at the extreme states 9.5 and 13.5. When HgB reaches a dangerously low level, the maximum EPO dose is the only feasible action. On the other hand, when HgB level is too high, EPO should be withheld. These extreme states are undesirable and it is expected that the system avoids them.

For safety of the patient, we limited the exploration only to time instances when the system was visiting the target state. In other words, ϵ is nonzero only when $11.0 \leq x_1 \leq 12.0$. In this case, the exploration probed how decreasing EPO affects the patient's response. Such an exploration aims at minimizing the patient exposition to the drug, as well as the total EPO administered.

In the simulation involving the Q-learning procedure, we picked λ as 0.1, the diminishing learning rate as $\nu=1/k$, the discount factor as $\gamma=0.9$, the exploration parameter as $\epsilon=0.3$ when the system encountered the target states and $\epsilon=0$, otherwise. The spreads of the RBF nodes were picked as $\sigma=0.5$.

Figs. 2 and 3 show the progress of anemia management for a selected representative 'normal responder'. The top plots in each figure depict the HgB trajectory obtained as a result of administering EPO as shown in the plots second from the top. As an indicator of convergence of the Q-learning process, the third plot from the top of Fig. 2 presents the maximum norm of the changes in the Q-table along the treatment. The bottom plot shown in Fig. 2 shows the policy evolution. Each curve in the policy plot represents an action for the corresponding state as marked to the right of the plot. By analyzing the HgB trajectories, it can be concluded that Q-learning achieves the therapeutic goal.

This observation is also confirmed in Tables 1 and 2, where the statistics of the simulation are presented in terms of the mean value and the 95% confidence interval calculated over 100 patients for the following outcome measures:

- mean HgB level over the treatment period,
- standard deviation of HgB over the treatment period,
- total EPO used during the treatment.

The simulation statistics presented in these tables for 'normal responders' show no significant clinical differences

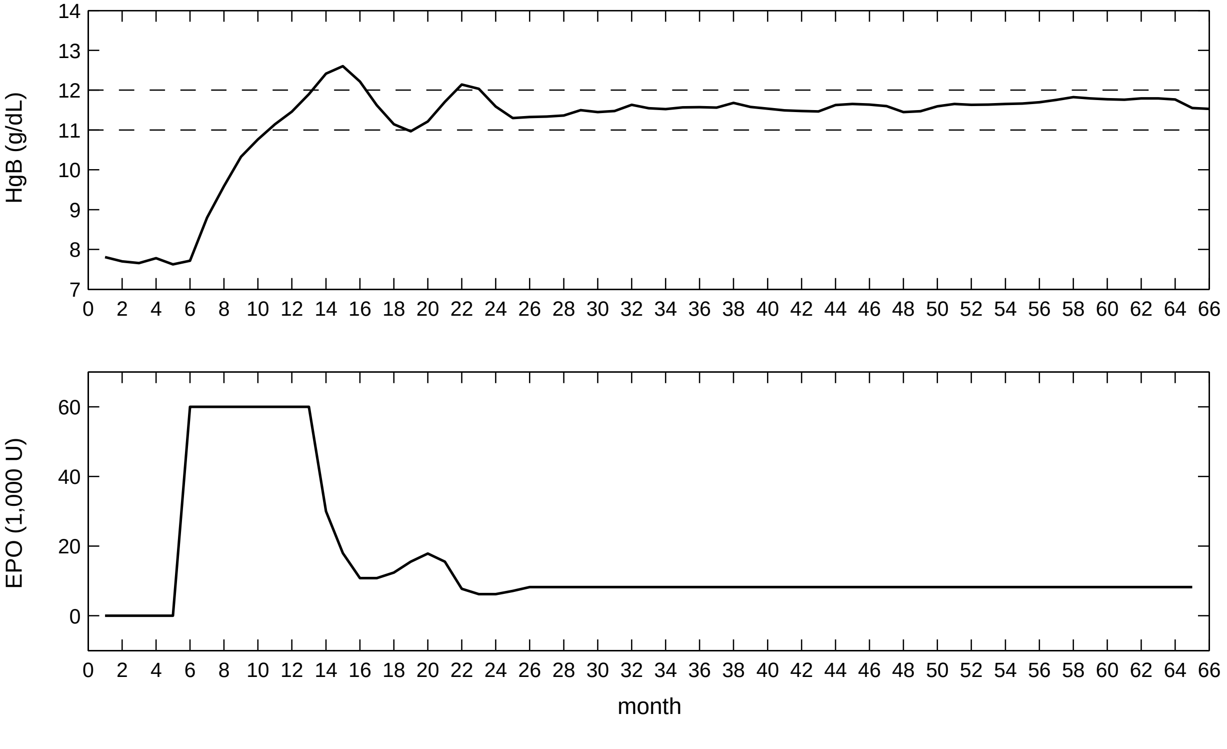

Fig. 5. HgB level (top) and EPO dose (middle) for an individual 'poor responder' as performed by AMP.

in terms of quality of anemia management between the two methods.

Figs. 4 and 5 show the progress of anemia management for a selected representative ‘poor responder’. The most profound difference between the HgB trajectory of a ‘poor responder’ and that of a ‘normal’ one is the time to get to the target range. It takes an average of 2 months for the HgB to get to the target range for a ‘normal responder’. For the ‘poor responder’, this period takes from 8 to 18 months. It can be observed that the *Q*-learning takes longer to get the HgB level of a ‘poor responder’ to the target range, compared to the AMP. This phenomenon can be attributed to the policy update occurring at the beginning of the therapy. Evidently, the initial action for HgB = 10.5 (30, 000 Units) is not aggressive enough for a ‘poor responder’ and causes a drop in HgB. As mentioned above, HgB below target is an undesired behavior, thus such an action receives a relatively low reward (or punishment) so that a different action is selected in the next step, based on the *Q*-table. This process contributes to increasing the time required to reach the target HgB range. Consequently, the AMP, as a prescribed policy, works faster for a ‘poor responder’, than *Q*-learning that ‘learns’ the policy on-the-fly. Nevertheless, statistics presented in Tables 1 and 2 for ‘poor responders’ show that the policy obtained by *Q*-learning and AMP achieve a comparable outcome.

5. Conclusions

In this work, an RL approach to individualized pharmacological management of anemia has been introduced. To enable numerical simulation of different types of patients, a Takagi-Sugeno fuzzy model was first built on basis of available patient data. We explored the implementation of *Q*-learning with RBF network for policy interpolation. Experimental evaluation allowed for comparing this method against the Anemia Management Protocol, regarded here as a ‘gold standard’. Statistical and clinical analysis of the test results showed that the *Q*-learning is capable of performing adequate anemia treatment in real time, comparable to the Anemia Management Protocol.

The research effort will now focus on using guided exploration in policy search. We will also investigate the use of adaptive sampling frequency of HgB. We expect that these two factors will further improve the quality of individualized anemia management.

Acknowledgements

The work of J. M. Zurada has been sponsored in part by the Systems Research Institute (IBS) of the Polish Academy of Science (PAN) 01-447 Warsaw, ul. Newelska 6. The work of M. E. Brier was funded by Department of Veteran Affairs, Merit Review Grant.

References

Bellman, R. (1983). *Mathematical methods in medicine*. Singapore: World Scientific Publishing.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.

Buell, J., Jelliffe, R., Kalaba, R., & Sridhar, R. (1970). Modern control theory and optimal drug regimens. II: Combination therapy. *Mathematical Biosciences*, *6*, 67–74.

Gaweda, A. E., Jacobs, A. A., Brier, M. E., & Zurada, J. M. (2003). Pharmacodynamic population analysis in chronic renal failure using artificial neural networks-a comparative study. *Neural Networks*, *16*, 841–845.

Gaweda, A.E., Muezzinoglu, M.K., Aronoff, G.R., Jacobs, A.A., Zurada, J.M., Brier, M.E. (2005). Reinforcement learning approach to individualization of chronic pharmacotherapy. Proceedings of the International Joint Conference on Neural Networks, July 31–August 4, 2005, Montreal, Canada.

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

Hu, C., Lovejoy, W. S., & Shafer, S. L. (1994b). An efficient strategy for dosage regimens. *Journal of Pharmacokinetics and Biopharmaceutics*, *22*, 73–92.

Moore, B.L., Sinzinger, E.D., Quasny, T.M., Pyeatt, L.D. (2004). Intelligent control of closed-loop sedation in simulated ICU patients. Proceedings of the 17th International Florida Artificial Intelligence Research Symposium Conference. Miami Beach, FL.

Sanner, R. M., & Slotine, J. J. M. (1992). Gaussian networks for direct adaptive control. *IEEE Transactions Neural Networks*, *3*, 837–863.

Schaeffer, A. J., Bailey, M. D., Shechter, S. M., & Roberts, M. S. (2004). Modeling medical treatment using Markov decision processes. In M. L. Brandeau, F. Sainfort, & W. P. Pierskalla (Eds.), *Handbook of operations research/management science applications in health care*. Boston, MA: Kluwer Academic Publishers.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Takagi, T., & Sugeno, M. (1985). *Fuzzy identification of systems and its applications to modelling and control IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15.

Tsitsiklis, J. N. (1994). Asynchronous stochatic approximation and *Q*-learning. *Machine Learning*, *16*, 185–202.

Watkins, C. I. C. H., & Dayan, P. (1992). *Q*-learning. *Machine Learning*, *8*, 279–292.