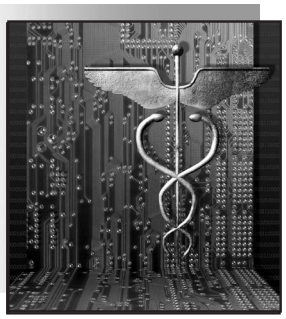


Using Clinical Information in Goal-Oriented Learning

Incorporating Clinical Expertise into Computational Framework to Improve Outcomes for Anemia Management

BY ADAM E. GAWEDA, MEHMET K. MUEZZINOGLU, GEORGE R. ARONOFF, ALFRED A. JACOBS, JACEK M. ZURADA, AND MICHAEL E. BRIER

Pharmacological treatment of chronic conditions often is a form of a recurrent trial and error. Typically, a physician administers a standard initial dose and observes the patient for a specific response and/or the occurrence of a side effect. Subsequently, the drug dose is adjusted in order to achieve a better response or to eliminate the dangerous side effect. This process continues until a desired response is obtained. The goal of this article is to provide and customize an efficient computational framework, based on reinforcement learning, to formalize this trial-and-error process.

Anemia due to end-stage renal disease (ESRD) is a common chronic condition in patients receiving hemodialysis [1]. It occurs due to an insufficient availability of a hormone called erythropoietin (EPO), which stimulates the production of red blood cells (erythropoiesis). The preferred treatment of renal anemia consists of external administration of recombinant human erythropoietin (rHuEPO). The Dialysis Outcomes Quality Initiative of the National Kidney Foundation recommends that the hemoglobin (Hgb) level in patients receiving rHuEPO be maintained between 11 and 12 g/dL. To follow these guidelines, dialysis units develop and maintain their own anemia management protocol (AMP).

Untreated, anemia can lead to a number of conditions including heart disease [2], [3], decreased quality of life [4], and increased mortality [5]. The sequencing and cloning of the human EPO gene and the subsequent availability of rHuEPO greatly improved morbidity and mortality for hemodialysis patients [6]–[8]. Prior to the clinical use of rHuEPO, chronic renal failure patients were the largest consumers of red blood cells. Ninety percent of dialysis patients require supplemental rHuEPO for the treatment of their anemia. In the United States, the cost of rHuEPO for treating these 320,000 dialysis patients exceeds US$1 billion annually [9].

Reinforcement learning is a computational approach that mimics a goal-oriented skill acquisition performed by humans and animals [10]. It represents the notion of goal-oriented learning by considering an agent; i.e., an intelligent decision system that interacts with an environment to achieve a specific goal through a trial-and-error process. Adopting this point of view for the drug administration problem, the agent represents the physician and the patient represents the environment. The process of administering the dose and observing the response corresponds to the trial-and-error aspect of learning. Applications of goal-oriented learning methods to drug administration have surfaced in the literature only very recently [11], [12]. We have recently proposed the use of two mainstream reinforcement learning methods to facilitate the individualization of anemia management. In [13] we have demonstrated that an on-policy reinforcement learning approach is capable of discovering appropriate dosing strategies for individuals with different responses to rHuEPO. On the other hand, in [14], we have shown that an off-policy reinforcement learning approach performs rHuEPO administration almost as well as the AMP that is currently used at our dialysis unit. These preliminary results made it possible for us to identify which aspects of the reinforcement learning algorithms can be further customized to better fit this specific application.

In this work, which is an extension of [15], we present an approach that incorporates prior knowledge about the dose-response characteristic of the patient into the learning. It is known that the dose-response curve of Hgb versus EPO has a monotonic shape [1]. For example, if the Hgb response is insufficient after a starting rHuEPO dose, the physician knows that the next dose should be higher. Consequently, an understanding of the processes that controls red blood cell production tells us in what direction a dose adjustment should take but not the size of the adjustment. The Q-learning algorithm, such as the one used in [14], is not able to utilize this type of prior knowledge. This may lead to suboptimal treatment outcomes.

Guiding reinforcement learning with external knowledge has been a major issue for over a decade. Many researchers have adopted the term *advice* to identify this type of knowledge provided/imposed by an external source. Two natural problems that arise when dealing with advice are how to represent it and where to incorporate it in the learning system. Maclin and Shavlik [16] proposed a reinforcement learning system that requests advice from an external observer and assimilates the provided information in its internal structure. On the other hand, a learning scheme augmented with an explicit supervisor is presented in [17]. This study addresses the issue of combining the supervisor knowledge with the reinforcement signal in an optimal

 0739-5175/07/$25.00©2007IEEE

Reinforcement learning is a computational framework that mimics trial-and-error learning performed in humans and animals.

way. These two works present and delineate very sound scenarios of utilizing advice, which is however provided on-the-fly, not a priori. In our problem, we consider another form of advice, which originates from clinical practice and gives rise to an efficient abstraction, simplifying the search for an optimal dosing policy. We propose a simple modification to Q-learning to allow for the use of prior information about the character of the Hgb versus EPO curve. We expect that this modification will make the rHuEPO dosing more efficient.

Understanding the Data

The proposed simulation scenario for the dosing of EPO is shown in Figure 1. The dynamics governing the behavior of Hgb are represented by the block denoted "Patient" and are influenced by two factors: the administered rHuEPO dose and available iron needed to make red blood cells measured by the transferrin saturation (TSat). The key element of the proposed method is represented by the block labeled "Q-learning Agent." This agent's task is to acquire the best dosing policy; i.e., the dose of rHuEPO for each possible Hgb level that will achieve the goal of the therapy. This policy is evolved during sequential observations of the Hgb changes in "Patient" as a result of changing rHuEPO dose. After an rHuEPO dose is administered to "Patient," a corresponding change of Hgb level occurs. The direction and the amount of this change are used by "Q-learning Agent" to improve the current policy. In this work, the policy is represented by a look-up table that relates the Hgb levels to the corresponding rHuEPO amounts; thus, it is a coarse policy. On the other hand, the Hgb levels observed in "Patient" are continuous. To allow for gradual dose changes, the "Dose Administration" block performs an interpolation to find an intermediate dose, should the actual Hgb level lie between the levels represented in the coarse policy.

The goal of the treatment is to drive the Hgb level in a patient to within the target range of 11 to 12 g/dL and maintain it within this range by adjusting the amount of rHuEPO administered. For the purposes of this work we chose the desired Hgb as the midpoint of this range, 11.5 g/dL. As noted above, the Hgb response is also influenced by the TSat, measured in percent, and represented in this work as a normal random variable with mean m_{TSat} and standard deviation σ_{TSat}. Translating this formulation into a state-space notation, frequently used in control theory, Hgb and TSat become state variables and rHuEPO becomes the control variable. In the proposed simulation scheme, "Patient" is represented by a stochastic iteration,

$$\mathrm{Hgb}[k+1] = f(\,\mathrm{Hgb}[k],\ \mathrm{rHuEPO}[k],\ \mathrm{TSat}[k]) \qquad (1)$$

In this formula, the index k denotes a time step, equal to 1 month, and $f(\,)$ represents a functional relationship. The details of function $f(\,)$ will be discussed below. Even though this equation is a fundamental component of the simulation and is specified a priori, it is unknown to the Q-Learning Agent. The function $f(\,)$ is used in this study exclusively for the purpose of representing the Patient module within the framework presented in Figure 1.

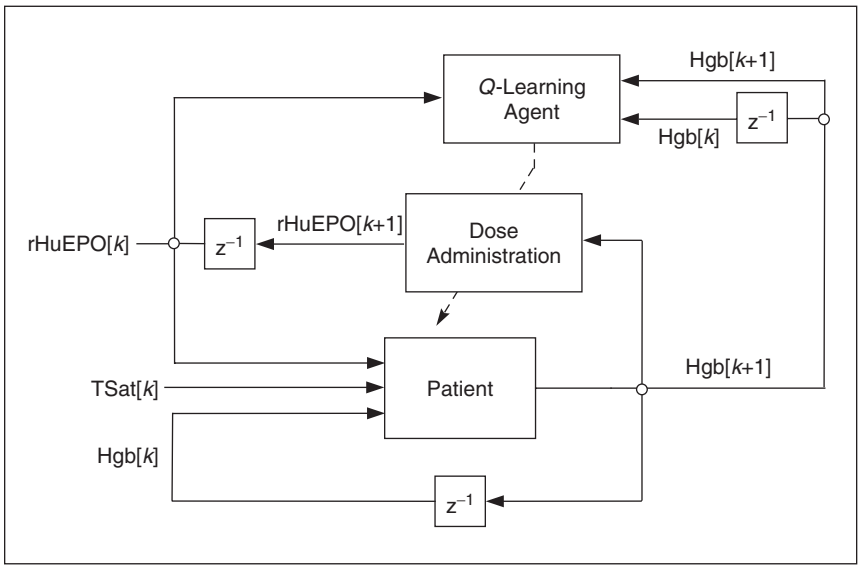


Fig. 1. Block diagram of the simulation scenario for reinforcement learning-based anemia management.

To simulate the Hgb response to rHuEPO, we developed a patient model based on a Takagi-Sugeno (TS) fuzzy system [18]. A TS fuzzy system can be regarded as a collection of local models, whose partial contribution toward the system output is determined by the system input. We use the patient's resistance to rHuEPO as the input. This quantity is computed as a ratio of the 3-month average rHuEPO dose to the 3-month average level of Hgb. We hypothesized that the Hgb level in individuals with low rHuEPO resistance was governed by different dynamics than those with high rHuEPO resistance. The TS fuzzy system was determined to be an effective tool to represent the imprecise boundary

between the notions of "low" and "high" rHuEPO resistance. To build the patient model, we used data from 186 patients undergoing hemodialysis at the Division of Nephrology of the Department of Medicine, University of Louisville. Using data records of 12 months containing monthly Hgb levels and rHuEPO doses, we calculated rHuEPO resistance for each individual. We then performed fuzzy c-means clustering [19] and produced fuzzy membership functions categorizing the two types of rHuEPO resistance. These membership functions are shown in Figure 2.

We will subsequently refer to individuals with low rHuEPO resistance as "normal responders" and to individuals with high rHuEPO resistance as "poor responders." However, a single individual usually exhibits characteristics that combine the features of both response types to a certain degree (μ). For example, someone with rHuEPO resistance of 1.0 is similar to a normal responder to a degree 0.80, and to a poor responder to a degree 0.28.

Having extracted the fuzzy classification of rHuEPO resistance, we created the models (1) for each response group. The data set used to estimate these models contained monthly Hgb, rHuEPO, and TSat values. We used the following linear model:

$$\begin{aligned}\mathrm{Hgb}[k+1] = {} & c_{\mathrm{Hgb}}\,\mathrm{Hgb}[k] + c_{rHuEPO}\,\mathrm{rHuEPO}[k] \\ & + c_{TSat}\,\mathrm{TSat}[k] + c_{\mathrm{bias}}. \end{aligned} \tag{2}$$

In this model, the coefficients $\mathbf{c} = [c_{Hgb}, c_{rHuEPO}, c_{TSat}, c_{\mathrm{bias}}]$ determine the Hgb response characteristic of that response group. We estimated these coefficients using the weighted least squares method and obtained the following values

$$\mathbf{c}^{\mathrm{high}} = [\,0.81 \quad 0.03 \quad -0.014 \quad 1.40\,] \tag{3}$$

for the high rHuEPO resistance group and

$$\mathbf{c}^{\mathrm{low}} = [\,0.72 \quad 0.05 \quad -0.006 \quad 3.21\,] \tag{4}$$

for the low rHuEPO resistance group.

Using the TS approach [18], the coefficients c for the individual shown in Figure 2 can now be simply computed in the following way:

$$\begin{aligned} c_{\mathrm{Hgb}} &= \frac{\mu_{\mathrm{low}} c_{\mathrm{Hgb}}^{\mathrm{low}} + \mu_{\mathrm{high}} c_{\mathrm{Hgb}}^{\mathrm{high}}}{\mu_{\mathrm{low}} + \mu_{\mathrm{high}}} \\ &= \frac{0.80 \cdot 0.72 + 0.28 \cdot 0.81}{0.80 + 0.28} = 0.74 \end{aligned} \tag{5}$$

$$\begin{aligned} c_{rHuEPO} &= \frac{\mu_{\mathrm{low}} c_{rHuEPO}^{\mathrm{low}} + \mu_{\mathrm{high}} c_{rHuEPO}^{\mathrm{high}}}{\mu_{\mathrm{low}} + \mu_{\mathrm{high}}} \\ &= \frac{0.80 \cdot 0.81 + 0.28 \cdot 0.72}{0.80 + 0.28} = 0.045 \end{aligned} \tag{6}$$

$$\begin{aligned} c_{TSat} &= \frac{\mu_{\mathrm{low}} c_{TSat}^{\mathrm{low}} + \mu_{\mathrm{high}} c_{TSat}^{\mathrm{high}}}{\mu_{\mathrm{low}} + \mu_{\mathrm{high}}} \\ &= \frac{0.80 \cdot (-0.006) + 0.28 \cdot (-0.014)}{0.80 + 0.28} \\ &= -0.008 \end{aligned} \tag{7}$$

$$\begin{aligned} c_{\mathrm{bias}} &= \frac{\mu_{\mathrm{low}} c_{\mathrm{bias}}^{\mathrm{low}} + \mu_{\mathrm{high}} c_{\mathrm{bias}}^{\mathrm{high}}}{\mu_{\mathrm{low}} + \mu_{\mathrm{high}}} \\ &= \frac{0.80 \cdot 3.21 + 0.28 \cdot 1.40}{0.80 + 0.28} = 2.74. \end{aligned} \tag{8}$$

To summarize, the final patient model is a nonlinear weighted combination of the two linear models that define the response for a normal and poor responder.

The Q-learning Agent is capable of learning the dosing strategy regardless of the model structure and its parameters. Therefore, other types of models may be used for this purpose as well [20]–[22]. The above-described model was used in this study because the fuzzy response classification provides a simple and convenient way of representing the characteristics of a patient in terms that are easily recognized in the medical field.

The goal of the agent can be summarized as learning the best dosing policy (Hgb – rHuEPO pairs) using a trial-and-error approach. In much the same way as the physician, the agent recommends a specific rHuEPO dose to be administered and assesses its effect on the Hgb level with respect to the therapeutic goal. The acquired information not only determines the adequacy of the recommended dose but also indicates how it should be modified to improve the response. More specifically, an increase or decrease of the dose amount may be recommended next time the same Hgb level is measured. The agent maintains a dosing policy in the form of a look-up table of current Hgb and recommended rHuEPO dose pairs that will move the Hgb level to the desired range. This look-up table is updated as the agent gains experience. An update to this policy occurs upon observing the change of patient's Hgb level, following administration of the recommended rHuEPO dose. It can be shown that, under some mild technical assumptions, this algorithm produces an optimal policy [23]; i.e., one that, for each Hgb level, selects the rHuEPO dose that is most beneficial in terms of a long-term treatment outcome. We provide the computational details of this learning scheme next.

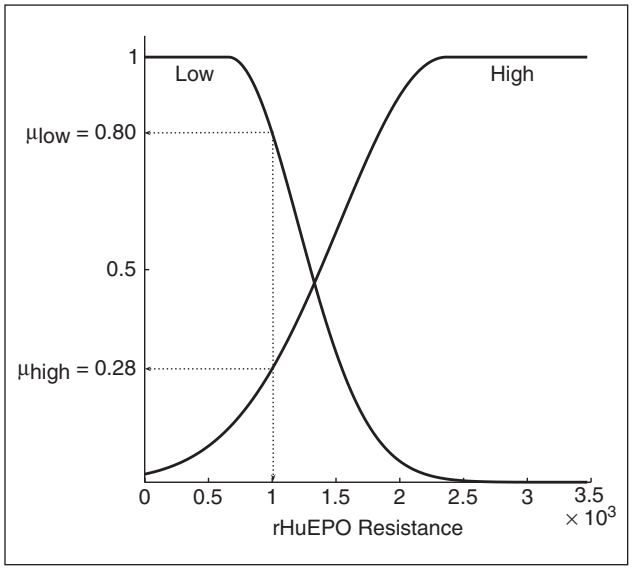


Fig. 2. Fuzzy membership functions for EPO resistance index.

In the proposed simulation scenario, the agent uses a policy defined on finite Hgb and rHuEPO ranges. In the data set of 186 patients used for this study, the Hgb ranged from 7 to 14 g/dL. Laboratory measurements specify Hgb values with accuracy of one decimal place. Hence, the smallest possible increment of the Hgb range would be every 0.1 g/dL. This, however, would greatly increase computation time and is not meaningful from a medical perspective since a step size of 0.1 g/dL is smaller than the measurement error for Hgb. In this work, we represent the Hgb space as a range from 7 to 14 g/dL with a step of 0.5 g/dL. The "Dose Administration" block in Figure 1 is responsible for computing intermediate doses for Hgb levels not represented within the policy. The rHuEPO doses in the data set ranged from 0 to 60,000 units per week. The smallest minimum dose per treatment is 1,000 Units and there are three treatments per week. Therefore, we represented the rHuEPO space as a range from 0 to 60,000 units with a step of 3,000 units.

The experience acquired by the agent is stored within a table computed using the following formula [8]:

$$\begin{aligned} Q(\text{Hgb, rHuEPO}) =& E\Big[\sum_{i=0}^{\infty}\gamma^i g(\text{Hgb}[k], \text{Hgb}[k+1]) \\ & |\text{Hgb}[0] \\ =& \text{Hgb, rHuEPO}[0] = \text{rHuEPO}\Big], \end{aligned} \quad (9)$$

which states that a Q-value for a Hgb/rHuEPO combination is an expected value of the sum of immediate rewards $g(\text{Hgb}[k], \text{Hgb}[k+1])$ for changing the Hgb level from Hgb$[k]$ to Hgb$[k+1]$, discounted by the coefficient $\gamma \in [0, 1]$. The reward function $g(\ )$ is the only component of the learning algorithm that links the agent to the control objective. As the treatment goal is to drive the Hgb level to the target interval of 11 to 12 g/dL, we used a $g(\ )$ function that reinforces all Hgb changes toward that interval and attenuates all Hgb changes away from it. Furthermore, since the observed Hgb is not precisely measured by the analytical technique and varies in time due to other unmeasured factors, we decided to give the strongest reinforcement to any Hgb changes toward the median of the target range. Maintaining the Hgb level close to 11.5 g/dL minimizes the likelihood of Hgb being outside the target range in a real patient. The reward function used in this work is summarized in equation (10), as shown at the bottom of the page.

The Q-value defined by (9) can be viewed as a measure of a long-term benefit from administering a certain dose of rHuEPO at a given Hgb level. In other words, for a set of possible Hgb/rHuEPO combinations, their Q-values simply denote how preferable it is to apply a specific rHuEPO dose at that Hgb level. The higher the Q-value, the more preferable the specific the rHuEPO dose is. We will further denote the most preferred dose by rHuEPO*. Once the table of all possible Q-values (Q-table) is available, rHuEPO* for each Hgb level can be found using a greedy policy search:

$$\text{rHuEPO}^*(\text{Hgb}) = \arg\max_{\text{Hgb}} Q(\text{Hgb, rHuEPO}). \quad (11)$$

Given a Hgb change from Hgb$[k]$ to Hgb$[k+1]$ as a result of administering specific dose rHuEPO$[k]$, together with the incurred reward g(Hgb$[k]$, Hgb$[k+1]$), the agent incrementally updates the Q-values [23]:

$$\begin{aligned} Q(\text{Hgb}[k], \text{rHuEPO}[k]) =& Q(\text{Hgb}[k], \text{rHuEPO}[k]) \\ & + \alpha(k)[g(\text{Hgb}[k], \text{Hgb}[k+1]) \\ & + \gamma \max_{\text{rHuEPO}} Q(\text{Hgb}[k+1], \text{rHuEPO}) \\ & - Q(\text{Hgb}[k], \text{rHuEPO}[k])], \end{aligned} \quad (12)$$

where $\alpha(k)$ is an exponentially decreasing learning rate. Since the formulation of the reward function (10) does not take into account delayed reinforcement (i.e., only current and not past Hgb changes are being taken into account), the discount coefficient γ does not influence the learning process.

In the standard formulation, the agent updates only the Q-values for Hgb/rHuEPO pairs that are "encountered" during the dosing process. In critical applications, such as this one, this may lead to long learning times, as we have observed in [13]. For the rHuEPO dosing problem described here, it has been known that the response of Hgb to rHuEPO is monotonic and nondecreasing [24]. We propose an extension to the standard Q-learning algorithm that updates the Q-values not only for the current Hgb/rHuEPO combination but also for those combinations that are deemed not preferable, based on the experience acquired by the agent and the assumption of monotonicity. As a result, those rHuEPO doses are less likely to be administered at a further stage, which in turn increases the learning speed. The multiple Q-value updates are performed according to the following heuristic rules:

- If Hgb$[k]$ $<$ 11.5 and Hgb$[k+1] \le$ Hgb$[k]$, or Hgb$[k] = 11.5$ and Hgb$[k+1] <$ Hgb$[k]$ then update Q(Hgb, rHuEPO) for all Hgb $\le$ Hgb$[k]$ and rHuEPO $\le$ rHuEPO$[k]$. In other words, if the current Hgb level is below 11.5 g/dL and the current rHuEPO does not increase or causes a decrease of the Hgb level, or the current Hgb level is 11.5 g/dL and the current rHuEPO causes a decrease of the Hgb level, all doses below and including the current one are inadequate for the current Hgb level and below.
- If Hgb$[k] >$ 11.5 and Hgb$[k+1] \ge$ Hgb$[k]$, or Hgb$[k] = 11.5$ and Hgb$[k+1] >$ Hgb$[k]$ then update Q(Hgb, rHuEPO) for all Hgb $\ge$ Hgb$[k]$ and rHuEPO $\ge$ rHuEPO$[k]$. In other words, if the current Hgb level is above 11.5 g/dL and the current rHuEPO does not decrease or causes an increase of the Hgb level, or the current Hgb level is 11.5 g/dL and the current rHuEPO causes an increase of the Hgb level, all doses above and including the current one are inadequate for the current Hgb level and above.
- Otherwise perform a standard Q-update.

$$g(\text{Hgb}[k], \text{Hgb}[k+1]) = \begin{cases} -1, & 11.5 \ge \text{Hgb}[k] > \text{Hgb}[k+1] \quad \vee \quad \text{Hgb}[k+1] > \text{Hgb}[k] \ge 11.5 \\ 0.5, & 11.5 > \text{Hgb}[k+1] > \text{Hgb}[k] \quad \vee \quad \text{Hgb}[k] > \text{Hgb}[k+1] > 11.5 \\ 1, & \text{Hgb}[k+1] = 11.5 \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Drug administration to achieve therapeutic effect and avoid adverse effects is often a trial-and-error process.

The proposed extension to the standard Q-learning algorithm performs multiple, region-based updates and can be viewed as a form of state abstraction, a general methodology that improves the learning speed. The earliest work on using region-based updates in reinforcement learning was by Yee et al. [25]. In this work, the definition of a region was based on a hierarchical organization of concepts. Dietterich et al. [26] used state abstraction as a component of their algorithms for planning in deterministic and stochastic domains. These methods assumed perfect knowledge about the environment and adaptively created regions to be evaluated during learning. Experiments verified that the region-based updates accelerate learning.

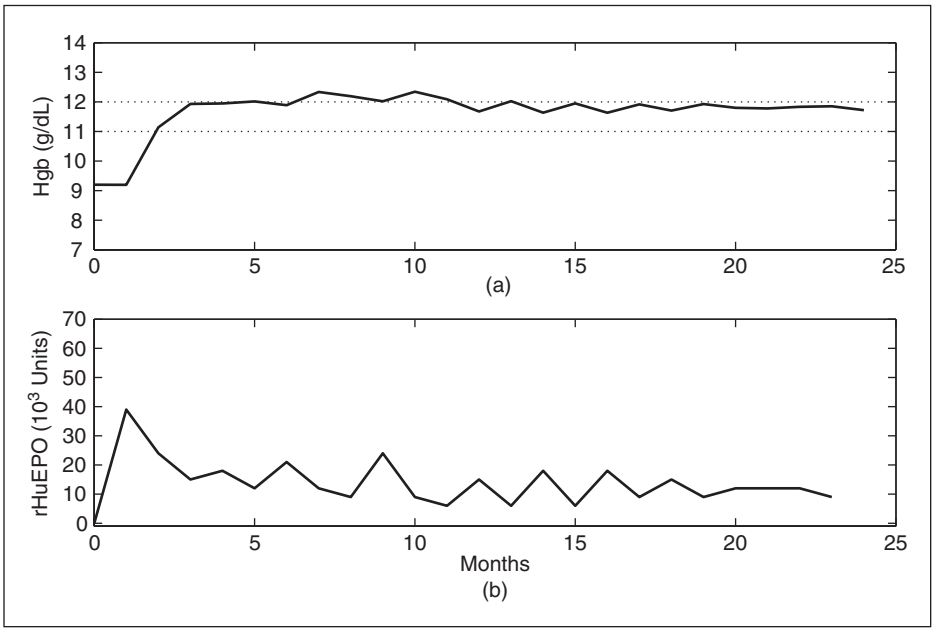


Fig. 3. Example of a simulated anemia treatment for a normal responder as performed by standard Q-learning: (a) plot of Hgb level; (b) plot of administered rHuEPO.

As the Q-learning Agent operates on a policy defined on finite Hgb levels, an interface is required to compute intermediate rHuEPO doses; i.e., doses corresponding to Hgb values not represented in the policy. This interface is represented by the Dose Administration block in Figure 1. If we denote a measured Hgb level in the patient by Hgb_p, then the corresponding intermediate dose, rHuEPO_p, is calculated by Dose Administration simply by performing linear interpolation

$$\text{rHuEPO}_p = \text{rHuEPO}_l + \frac{\text{Hgb}_u - \text{Hgb}_p}{\text{Hgb}_u - \text{Hgb}_l} \times (\text{rHuEPO}_u - \text{rHuEPO}_l), \qquad (13)$$

where the subscripts l and u denote the lower and upper nearest Hgb levels in the policy and their corresponding rHuEPO doses. For example, if Hgb_p is measured as 10.2 g/dL, then Hgbl is 10.0 g/dL and Hgb_u is 10.5 g/dL. Let us assume that, for a given policy, rHuEPO_l is 21,000 units and rHuEPO_u is 12,000 units. The calculated rHuEPO_p value will then equal 15,600 units, which, after rounding down to multiples of 3,000 units, will give the recommended rHuEPO dose of 15,000 units.

Preparation of the Data

We evaluated the proposed method using an experimental setup similar to the ones used in [13], [14], and [15]. We first created an artificial population of 100 normal responders and 100 poor responders. For each artificial patient, a set of four initial Hgb values was generated based on the statistics from the above-mentioned patient data from 186 individuals. Since our aim was to test incident patients (i.e., ones that had not received rHuEPO previously), we had to impose their rHuEPO resistance a priori. For normal responders, we randomly generated their degrees of membership in the *low* rHuEPO resistance group from the range 0.6 to 0.9. Conversely, their degree of membership in the *high* rHuEPO resistance group was randomly drawn from the range 0.1 to 0.4. Similar procedures were then performed for poor responders. For each patient, the treatment was simulated over a period of 24 months. Before each treatment simulation, we initialized the Q-table such that the first policy would represent a so called "best-guess" dosing strategy. This "best guess" strategy is based on a heuristic that low Hgb levels require large rHuEPO doses and vice versa. Following the explanation presented in the Understanding the Data section, we set the discount coefficient γ to 0. We set the initial value α_0 of the learning rate $\alpha(k)$ to 0.33 and decreased it by $1/k$ in each step. The mean value of TSat, m_{Tsat}, for each patient was randomly drawn from the interval 10 to 50% and the standard deviation of TSat, σ_{Tsat}, was set to 10%. These values were decided upon based on the analysis of our 186 real patients. Furthermore, the TSat variation during the simulated treatment was bounded between 10 – 50%.

Anemia management in renal failure is a good example for testing new techniques of drug administration.

In the first set of simulations, we applied basic Q-learning, i.e., one without multiple Q-updates. Subsequently, we repeated the simulations using the proposed extended Q-learning algorithm that includes multiple Q-updates using the same experimental conditions. We completed the evaluation process by simulating the anemia treatment with a numerical implementation of the AMP currently used at the Division of Nephrology to administer rHuEPO.

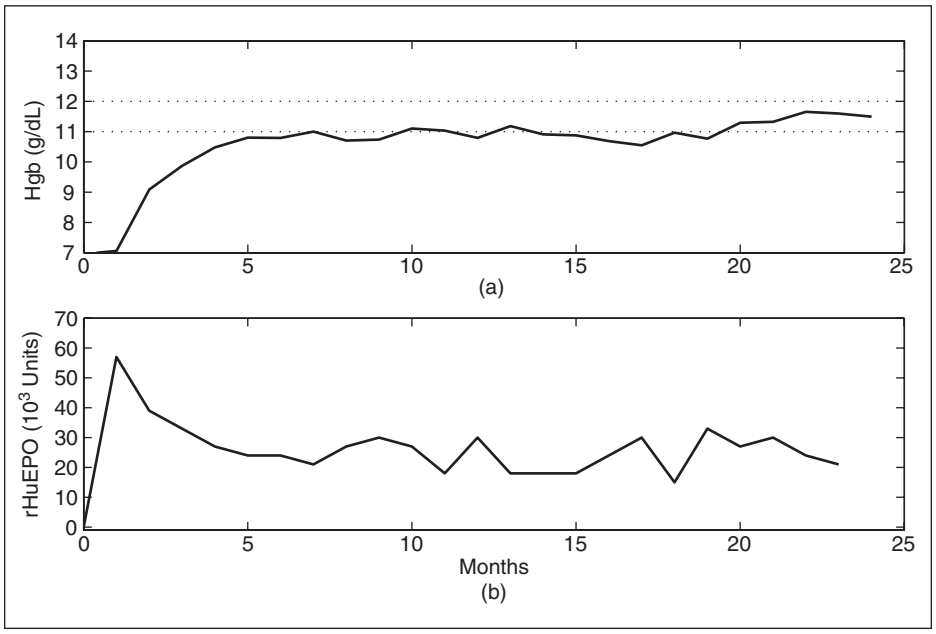


Fig. 4. Example of a simulated anemia treatment for a poor responder as performed by standard Q-learning: (a) plot of Hgb level; (b) plot of administered rHuEPO.

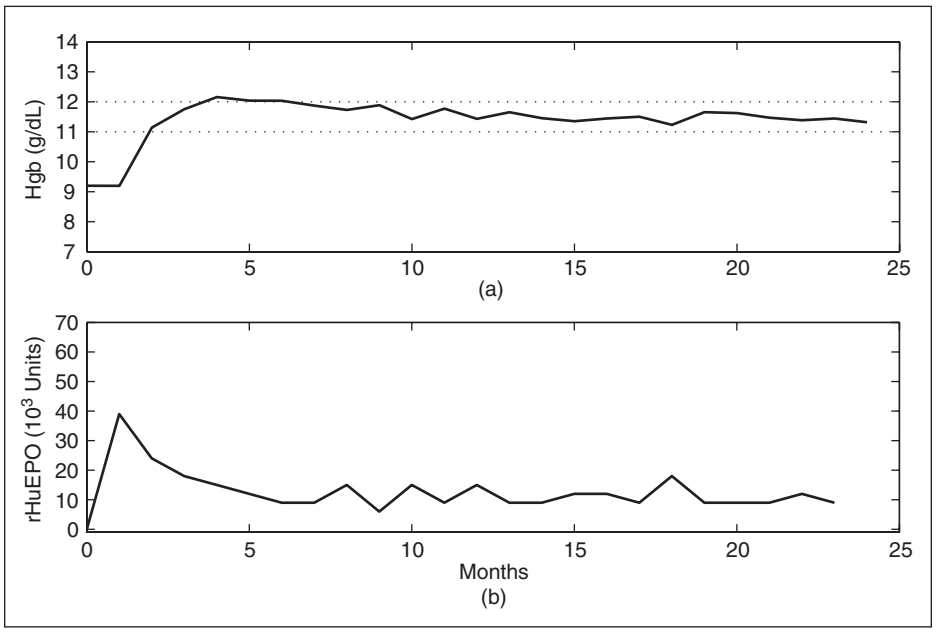


Fig. 5. Example of a simulated anemia treatment for a normal responder as performed by Q-learning with multiple updates: (a) plot of Hgb level; (b) plot of administered rHuEPO.

To allow for a comparative analysis between the tested methods, we used the following criteria:

- mean Hgb level—to assess the ability to drive the Hgb level to the target range
- standard deviation of Hgb—to determine how well a constant Hgb level is maintained
- number of times Hgb is out of target range—to determine how well the Hgb level is maintained within the target range; i.e., the clinical rate of success
- total rHuEPO used—to determine the cost-effectiveness.

Data Mining

Figures 3 and 4 show examples of a simulated anemia treatment performed by the standard Q-learning for a representative normal responder and poor responder, respectively. These figures demonstrate that the standard Q-learning exhibits a tendency to maintain the Hgb level close to the upper bound of the target range in normal responders and close to the lower bound of the target range in poor responders. Figures 5 and 6 show examples of a simulated anemia treatment performed by the extended Q-learning method with multiple Q-updates for a representative normal responder and a poor responder, respectively. These figure show that the

addition of multiple updates allows for much better control of Hgb level. For both, normal and poor responders, the Hgb level is now much closer to the median of the target range. Hence, one would expect this method to be more effective than the classical Q-learning in a real clinical environment. Finally, Figures 7 and 8 show examples of a simulated anemia treatment performed by the numerically implemented AMP. The most striking phenomenon that can be observed in these figures is the Hgb level fluctuation within the target range. This fluctuation occurs for both types of responders. This observation in the simulated environment is consistent with actual data from the clinical environment.

Tables 1 and 2 provide a quantitative comparison between the three simulated methods. The results are reported as means and 95% confidence intervals. Comparing the mean Hgb levels of normal responders between the three methods, we observe that Q-learning has a tendency to overcontrol the Hgb level, whereas Q-learning with multiple updates and the AMP drive the Hgb level to the target range. Comparing the standard deviations of the Hgb levels for normal responders among the three methods, one can observe that both Q-learning methods are more stable than the AMP. Due to the inability of Q-learning to maintain the Hgb level within the target range, the third criterion (number of times Hgb is out of target range) has a much larger value for this method compared to the other two. The amounts of administered rHuEPO are not significantly different between the three methods.

Comparing the mean Hgb levels of poor responders among the three methods, we can observe that all three methods are capable of driving the Hgb level to the target range. Comparison of the standard deviations of Hgb levels reveals that, similarly to the case of normal responders, Q-learning methods provide more stable Hgb control than the AMP. In terms of the number of times the Hgb level was out of target range, Q-learning with multiple updates outperformed the other two competitors. The amounts of administered rHuEPO are again not significantly different among the three methods.

Evaluation of Discovered Knowledge

Determining a protocol for the administration of a drug that results in adequate therapeutic concentrations and avoids toxic concentrations or adverse effects is a process that is often driven by trial and error and is time consuming. The use of rHuEPO for the treatment of the anemia of renal failure is a good model to test new techniques for the derivation of these drug dosing policies. We looked at two methods, standard Q-learning and Q-learning with multiple updates, to determine if we could improve on this process. We compared the results using simulation to a method used in clinical practice, the

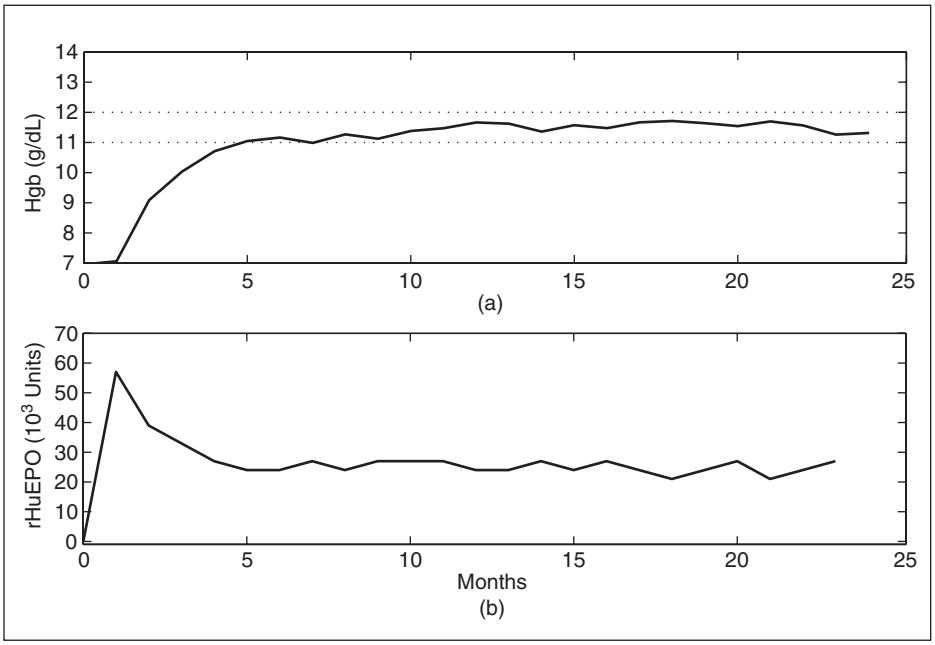


Fig. 6. Example of a simulated anemia treatment for a poor responder as performed by Q-learning with multiple updates: (a) plot of Hgb level; (b) plot of administered rHuEPO.

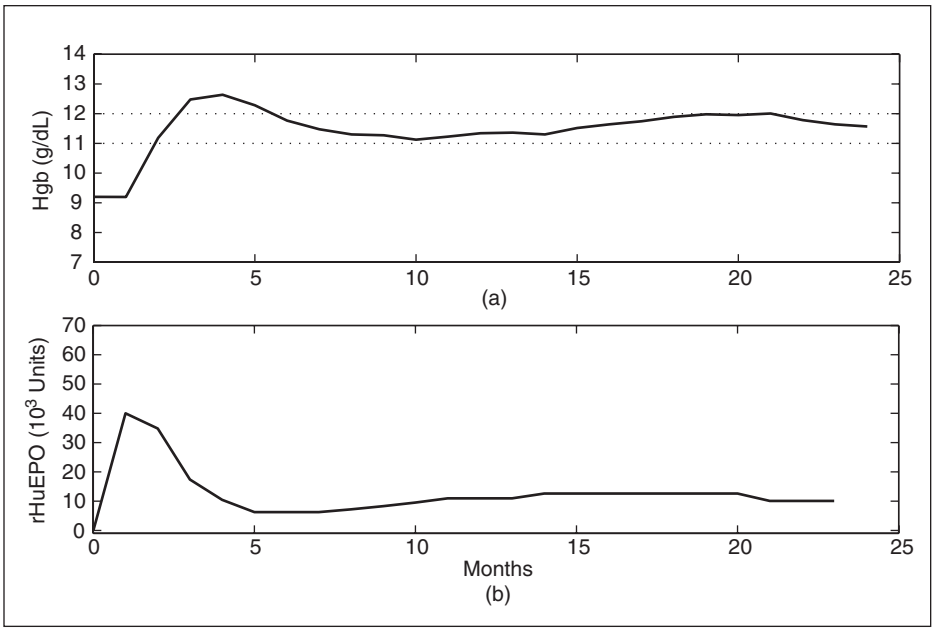


Fig. 7. Example of a simulated anemia treatment for a normal responder as performed by anemia management protocol: (a) plot of Hgb level, bottom; (b) plot of administered rHuEPO.

AMP, that has been developed over several years, Our data demonstrate the usefulness of the Q-learning methods as applied to drug dosing, where Q-learning with multiple updates preformed well based on the metrics that we used to judge between the compared methods.

We found that Q-learning is a useful tool to determine appropriate drug dosing schemes used to attain a desired drug concentration or response. In the case that we use to demonstrate these techniques, we are interested in the amount of Hgb in the blood. In general, as more EPO is given to a patient, the amount of Hgb in the blood will rise. However, there is a delay between the administration of the EPO and the change in Hgb level, and the process is further complicated by other factors, like the amount of iron available to make new Hgb. Therefore, the relationship is nonlinear in nature, it changes in time, and we cannot capture all of the variables that impact the Hgb variability. We have used the standard Q-learning in a previous application of dosing EPO and noticed several shortcomings that we attempted to address through the use of Q-learning with multiple updates. We had found previously that in this specific time-critical application, the standard Q-learning is too sensitive to the initial policy. Presented simulation results show that the standard Q-learning with multiple updates is not affected by this problem. More specifically, in our previous work, the standard Q-learning experienced trouble driving the Hgb level in poor responders into the target range. We were able to overcome this problem in our present work using a different, more aggressive initial policy and observing that, this time, the standard Q-learning struggled to drive the Hgb level down toward the target range in normal responders. From a clinical point of view, this is not as troublesome as patients with Hgb below the target range. Nevertheless, we observed that the Q-learning with multiple updates, being less sensitive to the initial policy, achieved adequate Hgb response in both normal and poor responders.

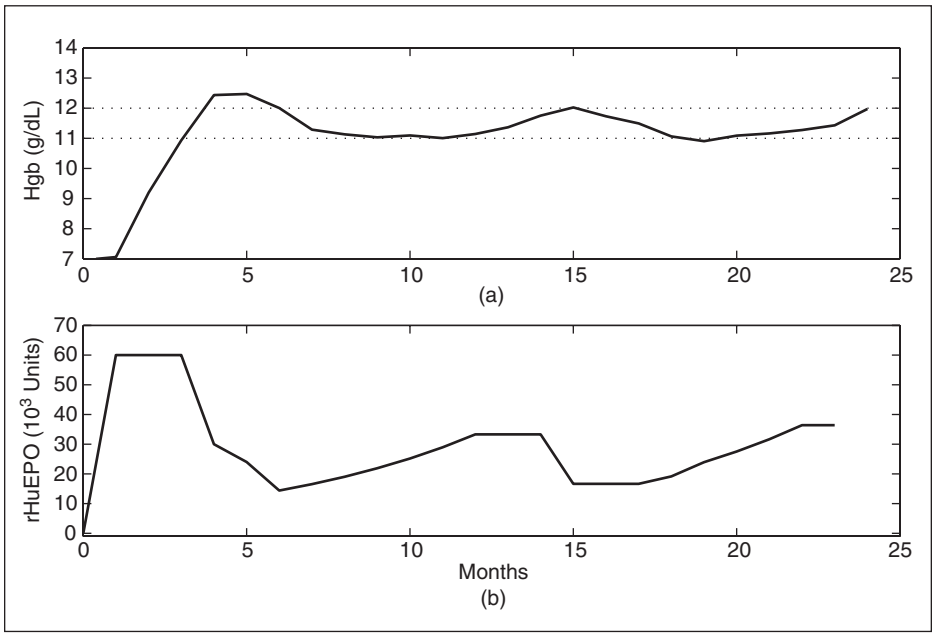


Fig. 8. Example of a simulated anemia treatment for a poor responder as performed by anemia management protocol: (a) plot of Hgb level; (b) plot of administered rHuEPO.

Looking at the results of the simulations from a clinical perspective, all three methods are fairly similar in achieving an average Hgb in the two subpopulations (poor and normal responders). The exception to this is that the standard Q-learning for normal responders results in an average Hgb of about 12.2 g/dL,

Table 1. Comparison of anemia treatment simulation results for normal responders.

Method	Standard Q-Learning	Q-Learning with Multiple Updates	Anemia Management Protocol
Mean Hgb (g/dL)	12.21 (11.77, 12.64)	11.77 (11.51, 12.02)	11.75 (11.47, 12.02)
Std Dev Hgb (g/dL)	0.16 (0.06, 0.26)	0.25 (0.09, 0.40)	0.50 (0.30, 0.70)
Number of Times Hgb Out of Target Range	15.4 (2.8, 28.0)	2.4 (0.0, 8.2)	2.9 (0.0, 6.6)
Total rHuEPO Used (1,000 Units)	286.3 (200.2, 372.5)	227.8 (122.3, 333.3)	223.2 (116.9, 329.6)

Table 2. Comparison of anemia treatment simulation results for poor responders.

Method	Standard Q-Learning	Q-Learning with Multiple Updates	Anemia Management Protocol
Mean Hgb (g/dL)	11.44 (10.96, 11.91)	11.46 (11.18, 11.73)	11.56 (11.34, 11.77)
Std Dev Hgb (g/dL)	0.26 (0.14, 0.37)	0.26 (0.12, 0.39)	0.58 (0.28, 0.87)
Number of Times Hgb Out of Target Range	1.3 (0.0, 7.7)	0.2 (0.0, 0.3)	2.9 (0.0, 11.0)
Total rHuEPO Used (1,000 Units)	468.1 (351.2, 585.1)	469.7 (334.3, 605.3)	474.9 (351.2, 585.1)

We found that *Q*-learning is a useful tool to help determine appropriate drug dosing stategies for achieving desired drug response.

which is outside the target range and is unacceptable given that the simulation represents a well-behaved environment without the random noise associated with real clinical data. *Q*-learning with multiple updates performs as well as the benchmark AMP and is slightly better in poor responders. The real difference then lies in the approximately 50% decrease in Hgb variability seen in the *Q*-learning with multiple updates, which is clinically significant. Currently, only about one-third of the patients treated with rHuEPO are within the recommended range of 11–12 g/dL Hgb. Decreasing the variability in mean Hgb will increase this percentage, which is a desirable effect.

The cost of EPO to Medicare and Medicaid in 1998 was about US$800 million per year [27]. We compared the total EPO utilization between the *Q*-learning and the AMP to verify whether the former would impact the amount of EPO used, as we have seen in other simulations using machine learning techniques [28]. This was not the case for either of the *Q*-learning methods. However, our formulation of the treatment goal in this study did not factor in the EPO utilization. This point will be of particular interest in further investigation to determine how the reinforcement learning techniques can be applied to enable not only an adequate but also cost-effective treatment of renal anemia.

Summary and Conclusions

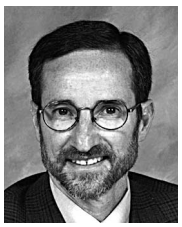
We have proposed an extension to the *Q*-learning algorithm that incorporates the existing clinical expertise into the trial-and-error process of acquiring an appropriate administration strategy of rHuEPO to patients with anemia due to ESRD. The specific modification lies in multiple updates of the *Q*-values for several dose/response combinations during a single learning event. This in turn decreases the risk of administering doses that are inadequate in certain situations and thus increases the speed of the learning process. We have evaluated the proposed method using a simulation test-bed involving an "artificial patient" and compared the outcomes to those obtained by a classical *Q*-learning and a numerical implementation of a clinically used administration protocol for anemia management. The outcomes of the simulated treatments demonstrate that the proposed method is a more effective tool than the traditional *Q*-learning. Furthermore, we have observed that it has a potential to provide even more stable anemia management than the AMP.

Acknowledgments

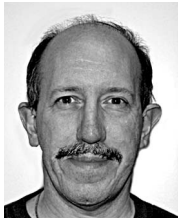
Portions of this work were supported by a grant from the Department of Veterans Affairs Merit Review Committee for Clinical Science Research and Development. The authors are thankful to anonymous reviewers for very constructive and inspiring comments.

Adam E. Gaweda received the M.Sc. in electrical engineering from Czestochowa University of Technology, Poland, and the Ph.D. in computer science and engineering from University of Louisville, Louisville, Kentucky, in 1997 and 2002, respectively. In 2002 he joined the Department of Medicine, Division of Nephrology, University of Louisville, where he currently holds the position of assistant professor. His research interests focus on application of computational intelligence and adaptive control to pharmacokinetic/pharmacodynamic modeling and drug administration.

Mehmet K. Muezzinoglu received the B.S. and M.S. from Istanbul Technical University, Istanbul, Turkey, in 1998 and 2000, respectively, and the Ph.D. from Dokuz Eylul University, Izmir, Turkey, in 2003. He has been pursuing his studies on recurrent neural networks and decision making under uncertainty at the Computational Intelligence Laboratory, University of Louisville, Louisville, Kentucky, since 2003, where he is also an assistant professor. Dr. Muezzinoglu was awarded a Ph.D. scholarship by the Scientific and Technical Research Council of Turkey (TUBITAK) Munir Birsel Foundation.



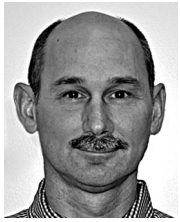
George R. Aronoff completed his internal medicine and nephrology training at the Indiana University School of Medicine. He received a M.S. in pharmacology at that institution and spent 2 years as a clinical pharmacology Fellow studying pharmacokinetics at Eli Lilly and Company. His research interests include the clinical pharmacology of drugs in patients with renal insufficiency. He is interested in applying novel biostatistical tools, including artificial intelligence and population kinetics to describing drug disposition and dosing in patients with impaired kidney function. Dr. Aronoff is the author or coauthor of over 100 published research manuscripts and is the editor of *Drug Prescribing in Renal Failure: Dosing Guidelines for Adults*, published by the American College of Physicians and now in its fourth edition.



Alfred A. Jacobs is a graduate of the University of Louisville School of Medicine, Louisville, Kentucky. He performed his residency in medicine and nephrology fellowship also at the University of Louisville. He received his Ph.D. in pharmacology and toxicology in

2004. Dr. Jacobs is the medical director of the in-center hemodialysis unit of the University of Louisville and his research interest are related to improving the hemodialysis treatment for these patients and in the application of artificial intelligence and intelligent control to patients with renal disease and those patients on hemodialysis.

Jacek M. Zurada is the Samuel T. Fife Alumni Professor and acting chairman of the Electrical and Computer Engineering Department at the University of Louisville, Louisville, Kentucky. He was the coeditor of *Knowledge-Based Neurocomputing* (MIT Press, 2000), *Computational Intelligence: Imitating Life* (IEEE Press), and *Introduction to Artificial Neural Systems* (PWS-Kent, 1992), and was a contributor to *Progress in Neural Networks* (Ablex, 1995). He is the author or coauthor of more than 240 journal and conference papers in the area of neural networks, data mining, image processing, and VLSI circuits. He is an associate editor of *Neurocomputing*. Dr. Zurada was an associate editor of the *IEEE Transactions on Circuits and Systems*. From 2001 to 2003, he was a member of the editorial board of the *Proceedings of the IEEE*. From 1998 to 2003, he was the editor-in-chief of the *IEEE Transactions on Neural Networks*. He has received a number of awards for distinction in research and teaching, including the 1993 Presidential Award for Research, Scholarship and Creative Activity. In 2001, he received the University of Louisville President's Distinguished Service Award for Service to the Profession. He has delivered numerous invited plenary conference presentations and seminars throughout the world. In 2003, he was conferred the title of the National Professor by the President of Poland and the Honorary Professorship of Hebei University, China. In 2004–2005, he was the president of the IEEE Computational Intelligence Society, of which he is also a Distinguished Speaker.



Michael E. Brier, Ph.D. is a graduate of Purdue University School of Pharmacy and Pharmacal Sciences in 1986 with a specialization in pharmacokinetics. He joined the faculty of the University of Louisville in 1987 as an Assistant Professor in Medicine where he continued his work in pharmacokinetics with an emphasis on population pharmacokinetics. His research in the application of intelligent dosing was first published in 1995 and focused on the use of an artificial neural network to predict drug levels in patients receiving an aminoglycoside antibiotic. His current research is funded by the Department of Veterans Affairs centering on the application of intelligent control to chronically administered drugs.

Address for Correspondence: Adam E. Gaweda,University of Louisville, Louisville, KY, 40292, USA. E-mail: agaweda@louisville.edu.

References

[1] J.W. Eschbach and J.W. Adamson, "Anemia of end-stage renal disease (ESRD)," *Kidney Int*., vol. 28, no. 1, pp. 1–5, 1985.
[2] J.D. Harnett, R.N. Foley, G.M. Kent, P.E. Barre, D. Murray, and P.S. Parfrey, "Congestive heart failure in dialysis patients: Prevalence, incidence, prognosis and risk factors," *Kidney Int*., vol. 47, no. 3, pp. 884–890, 1995.
[3] I.C. Macdougall, N.P. Lewis, M.J. Saunders, D.L. Cochlin, M.E. Davies, R.D. Hutton, K.A.A. Fox, G.A. Coles, and J.D. Williams, "Long-term cardiorespiratory effects of amelioration of renal anemia by erythropoietin," *Lancet*, vol. 335, no. 8688, pp. 489–493, 1990.
[4] D.L. Wolcott, J.T. Marsh, A. La Rue, C. Carr, and A.R. Nissenson, "Recombinant human erythropoietin treatment may improve quality of life and cognitive function in chronic hemodialysis patients," *Am. J. Kidney Dis*., vol. 14, no. 6, pp. 478–485, 1989.
[5] E.G. Lowrie, J. Ling, N.L. Lew, and Y. Yiu, "The relative contribution of measured variables to death risk among hemodialysis patients," in *Death on Hemodialysis: Preventable of Inevitable*?, E.A. Friedman, Ed. Boston, MA: Kluwer, pp. 121–141, 1994.
[6] J. Mocks, W. Franke, B. Ehmer, P. Scigalla, and O. Quarder, "Analysis of safety database for long-term rHuEPOetin-beta treatment. A meta-analysis covering 3697 patients." In *Pathogenetic and Therapeutic Aspects of Chronic Renal Failure*, K.M. Koch, G. Stein, Eds. New York: Marcel Dekker, pp. 163–179, 1997.
[7] Consensus Development Conference Panel, "Morbidity and mortality of renal dialysis: An NIH consensus conference statement," *Ann. Intern. Med*., vol. 121, no. 1, pp. 62–70, 1994.
[8] P. Grutzmacher, E. Scheuermann, I. Low, M. Bergmann, K. Rauber, R. Baum, J. Heuser, and W. Schoeppe, "Correction of renal anaemia by recombinant human erythropoietin: Effects on myocardial function," *Contrib. Nephrol*., vol. 66, pp. 176–184, 1988.
[9] U.S. Renal Data System, "USRDS 2001 annual data RrHuEPOrt: Atlas of end-stage renal disease in the United States," National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2001.
[10] P. Dayan and B.W. Balleine, "Reward, motivation and reinforcement learning," *Neuron*, vol. 36, no. 2, pp. 285–298, 2002.
[11] B.L. Moore, E.D. Sinzinger, L.D. Pyeatt, and T.M. Quasny, "Intelligent control of closed-loop sedation in simulated ICU patients," in *Proc. 17th Int. Florida Artificial Intelligence Research Society Conf*., Miami Beach, FL, 2004, pp. 109-113.
[12] S.A. Murphy, "An experimental design for the development of adaptive treatment strategies," *Statistics in Med*., vol. 24, no. 10, pp. 1455–1481, 2005.
[13] A.E. Gaweda, M.K. Muezzinoglu, G.R. Aronoff, A.A. Jacobs, J.M. Zurada, and M.E. Brier, "Reinforcement learning approach to chronic pharmacotherapy," in *Proc. 2005 Int. Joint Conf. Neural Networks,* Montreal, Canada, 2005, pp.3290-3295.
[14] A.E. Gaweda, M.K. Muezzinoglu, G.R. Aronoff, A.A. Jacobs, J.M. Zurada, and M.E. Brier, "Individualization of pharmacological anemia management using reinforcement learning," *Neural Netw*., vol. 18, no. 5-6, pp. 826–834, 2005.
[15] A.E. Gaweda, M.K. Muezzinoglu, G.R. Aronoff, A.A. Jacobs, J.M. Zurada, and M.E. Brier, "Incorporating prior knowledge into Q-learning for drug delivery individualization," in *Proc. 2005 Int. Conf. Machine Learning and Applications*, Los Angeles, CA, 2005, pp. 207–212.
[16] R. Maclin and J.W. Shavlik, "Creating advice-taking reinforcement learners," *Mach. Learn*., vol. 22, no. 1-3, pp. 251–281, 1996.
[17] M.T. Rosenstein and A.G. Barto, "Supervised actor-critic reinforcement learning," in *Handbook of Learning and Approximate Dynamic Programming*, J. Si, A. Barto, W. Powell, and D. Wunsch, Eds. pp. 359–380, New York: Wiley, 2004,
[18] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its application to modeling and control," *IEEE Trans. Syst., Man, Cybern*., vol. 15, Jan. Feb., pp. 116–132, 1985.
[19] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.
[20] R. Bellazzi, "Drug delivery optimization through Bayesian Networks: An application to erythropoietin therapy in uremic anemia," *Comput. Biomed. Res*., vol. 26, no. 3, pp. 274–293, 1992.
[21] J.D.M. Guerrero, E.S. Olivas, G. Camps-Valls, A.J. Serrano Lopez, J.J. Perez Ruixo, and N.V. Jimenez Torres, "Use of neural networks for dosage individualization of erythropoietin in patients with secondary anemia to chronic renal failure," *Comput. Biol. Med*., vol. 33, no. 4, pp. 361–373, 2003.
[22] A.E. Gaweda, A.A. Jacobs, M.E. Brier, and J.M. Zurada, "Pharmacodynamic population analysis in chronic renal failure using artificial neural networks: A comparative study," *Neural Netw*., vol. 16, no. 5–6, pp. 841–845, 2003.
[23] C. Watkins, "*Learning from delayed rewards*," M.S. thesis, Univ. of Cambridge, Cambridge, UK, 1989.
[24] J.F. Eliason, G. Van Zant, and E. Goldwasser, "The relationship of hemoglobin synthesis to erythroid colony and burst formation," *Blood*, vol. 53, no. 5, pp. 935–945, 1979.
[25] R.C. Yee, S. Saxena, P.E. Utgoff, and A.G. Barto, "Explaining temporal differences to create useful concepts for evaluating states," in *Proc. 8th National Conf. on Artificial Intelligence*, Boston, MA, 1990, pp. 882-888.
[26] T.G. Dietterich and N.S. Flann, "Explanation-based learning and reinforcement learning: A unified view," *Mach. Learn*., vol. 28, pp. 169–210, 1997.
[27] J.W. Greer, R.A. Milam, and P.W. Eggers, "Trends in use, cost, and outcomes of human recombinant erythropoietin,1989–98," *Health Care Financing Rev*., vol. 20, no. 3, pp. 55–62, 1999.
[28] A.E. Gaweda, A.A. Jacobs, G.R. Aronoff, and M.E. Brier, "Intelligent control for drug delivery in management of renal anemia," in *Proc. 2004 Int. Conf. Machine Learning and Applications*, Louisville, KY, 2004, pp. 355–359.