Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Model Predictive Control with Reinforcement Learning for Drug Delivery in Renal Anemia Management

Adam E. Gaweda, *Member, IEEE,* Mehmet K. Muezzinoglu, *Member, IEEE,* Alfred A. Jacobs, George R. Aronoff, and Michael E. Brier

Abstract — **Treatment of chronic conditions often creates the challenge of an adequate drug administration. The intra- and inter-individual variability of drug response requires periodic adjustments of the dosing protocols. We describe a method, combining Model Predictive Control for simulation of patient response and Reinforcement Learning for estimation of dosing strategy, to facilitate the management of anemia due to kidney failure.**

I. Introduction

Therapeutic drug delivery has long been recognized as a control problem [1]. In contrast to engineering fields, where the application of control methods has been quite successful, the nature of the drug dosing problem poses different challenges. The complexity of human body makes the development of an accurate mathematical model of the patient very difficult. In spite of this difficulty, the use of advanced control algorithms for drug delivery has been reported in the literature [2], [3], [4]. In [5], we simulated the use of Artificial Neural Network-based Direct Adaptive and Model Predictive Control for the management of renal anemia. We demonstrated that Model Predictive Control could improve the quality of anemia management, compared to currently used methods.

In the clinical practice treatment of chronic conditions often has a form of a recurrent trial and error process. Typically, a standard initial dose is administered first and the patient is observed for a specific response. The drug dose is then adjusted in order to improve the response, or to eliminate a potential side effect. To emulate this process numerically, we simulated the anemia management using Reinforcement Learning methods, such as SARSA [6] and Q-learning [7]. We demonstrated that these algorithms were capable of estimating adequate dosing strategies in the simulated environment.

Building upon the insights gained through [5] and [6], we present a combination of Model Predictive Control approach with Reinforcement Learning to establish a computer-based system for decision support in chronic drug dosing. We introduce the reader to the problem of anemia management first. Next, we provide a refresher of Model Predictive Control and Reinforcement Learning, and describe the proposed approach. Subsequently, we show the results of a simulated anemia management to demonstrate the feasibility of the proposed method. A summary and a discussion of the obtained results conclude the paper.

This work was supported in part by the U.S. Department of Veteran Affairs Merit Review Grant.

A. E. Gaweda , A. A. Jacobs, and G. R. Aronoff are with the Department of Medicine, Division of Nephrology, University of Louisville, Louisville, KY 40292, USA, 502-852-5757; e-mail: agaweda@kdp.louisville.edu.

M. K. Muezzinoglu is with the Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292 USA. (e-mail: mkerem@ieee.org).

M. E. Brier is with the Department of Veteran Affairs, Louisville, KY 40207, (e-mail: mbrier@kdp.louisville.edu).

II. Methods

A. Anemia Management

Anemia due to End Stage Renal Disease (ESRD) is a common chronic condition in hemodialysis patients [8]. It occurs due to an insufficient availability of a hormone called erythropoietin (EPO), which stimulates the production of red blood cells (erythropoiesis). Untreated, anemia can lead to a number of conditions including heart disease, decreased quality of life, and increased mortality. The preferred treatment of renal anemia consists of external administration of recombinant human erythropoietin (rHuEPO). The availability of (rHuEPO) greatly improved morbidity and mortality for hemodialysis patients. Ninety percent of hemodialysis patients require rHuEPO for the treatment of their anemia. In the United States, the cost of rHuEPO for treating these 320,000 dialysis patients exceeds $1 billion annually [9]. The Dialysis Outcomes Quality Initiative of 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 Protocols (AMP).

The Anemia Management Protocols are developed and updated based on a population response. Achieving a desired outcome in an individual is complicated due to variability of response within patient populations, and concurrent medications and comorbidities, specific for each patient. Frequently, the dose recommendation provided by the protocol is adjusted based on physician's experience and intuition. All this makes the anemia management very labor intensive. We test the hypothesis that a computer-based decision support tool will simplify this process, while achieving a comparable or better treatment outcome.

1-4244-0033-3/06/$20.00 ©2006 IEEE.

B. Model Predictive Control

A schematic diagram of the Model Predictive Controller (MPC) as applied to drug dosing is shown Figure 1. The controller contains two components, a predictive model of the patient and an optimizer for drug dose selection. The fundamental idea behind the MPC is to minimize a cost function related to the control goal [11], for example:

$$J = \sum_{k=1}^{H_p} \left(y_d(k) - y_m(k) \right)^2 . \quad (1)$$

This example represents the cost function J as a sum of squared differences between desired responses y_d and the responses predicted by the model y_m, over a time period H_p, after administering a dose (sequence) u'. The dose that minimizes the function J is administered to the patient and the process is repeated at the next dosing interval.

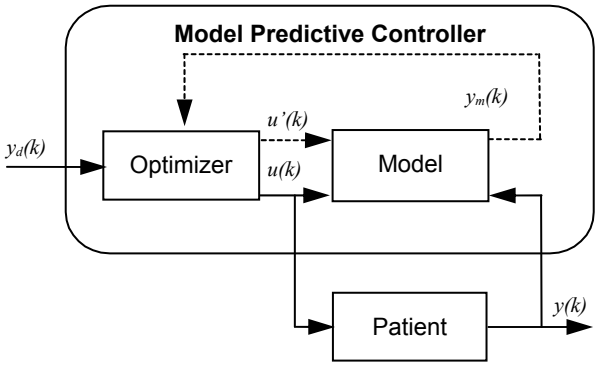

Fig. 1. Schematic diagram of Model Predictive Controller as applied to drug dosing. The symbol $y(k)$ represents patient's response, $y_m(k)$ – predicted response, $y_d(k)$ – desired response, $u(k)$ – recommended drug dose, $u'(k)$ – tested drug dose (sequence), and k is a time index.

The MPC approach requires the availability of a patient model. In this work, we use an Artificial Neural Network approach, as described in [10]. The most useful feature of the MPC with respect to the application to drug dosing is its ability to handle nonlinear control problems with time delays.

C. Reinforcement Learning

Reinforcement Learning (RL) is a collection of methods for producing optimal decision strategies through a process of trial and error [12]. In RL, learning is performed by an agent whose actions affect its environment. Based on the long-term learning objective, each action of the agent is either rewarded or punished. The optimal strategy is estimated from past actions and their consequences. RL is a promising technique for adaptive control problems where learning and control are performed simultaneously. Figure 2 shows a block diagram of an RL algorithm applied to the drug dosing problem, as described in [6] and [7]. The operation of this algorithm is centered around the so-called Q-table, which stores degrees of preferability for each action at a given state, i.e. state-action values, Q. Each action, i.e. drug dose, is followed by the observed response. This response is evaluated with respect to the long-term goal. For example, if the drug dose produced a favorable response, it would be rewarded (reinforced). On the other hand, if the observed response was undesired, the drug dose would be penalized. This feedback modifies relevant entries in the Q-table. The dosing policy, which specifies how to administer the drug based on patient response, is estimated by extracting the most preferable dose-response combinations from the Q-table.

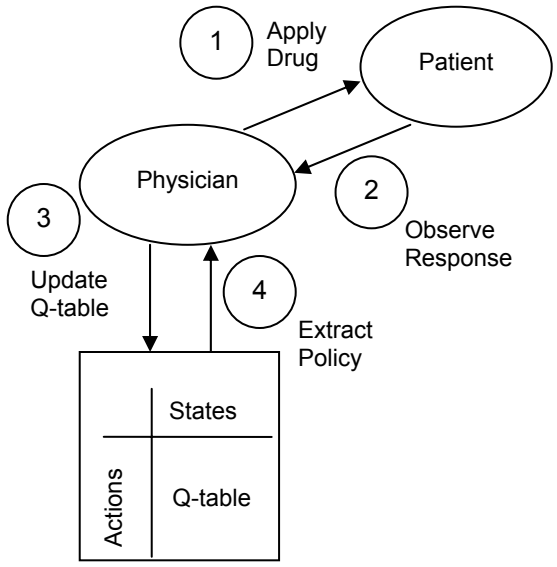

Fig. 2. Schematic diagram of Reinforcement Learning as applied to drug dosing. The Physician (agent) applies drug dose (action) to the Patient (environment) and observes Response (state). This information is used to update the Q-table and extract the dosing policy.

D. Model Predictive Control with Reinforcement Learning for Drug Delivery

The optimal control, u, in the MPC is found through an optimization process. When a linear model is used to predict the response, y_m, the optimal control u can be derived analytically. However, nonlinear models (such as Artificial Neural Networks) require much more involved optimization methods. Possible solutions involve Dynamic Programming, calculus of variations, optimization of a parametric control representation, and exhaustive search methods. The use of RL methods in MPC has recently been proposed in [13]. The authors formulated the MPC in terms of a Markov Decision Process and defined the cost function as a sum of state values, updated by the Temporal Difference method, TD(λ) [12].

Building upon the theoretical considerations presented in [13], we now present an approach that combines the most important features of the MPC and RL methods. We first described the use of the RL methods for drug delivery in [6]. We simulated estimation of drug dosing strategy using the on-policy RL method, SARSA [12]. In [7], we investigated the use of the off-policy method, Q-learning [14], in a simulated real time anemia management. We found that both

methods delivered performance comparable to that of a simulated Anemia Management Protocol. In this paper, we will use the on-policy method, SARSA, as the optimization mechanism for the drug dose determination in Model Predictive Control applied to management of renal anemia.

In SARSA, the learning progresses along an episode by observing state transitions and immediate reinforcements, and by adjusting the corresponding state-action values. If an action *a*, taken in state *s*, results in a transition to state *s'*, the next action being *a'*, then the quantity

$$\delta = r + \gamma\, Q(s',a') - Q(s,a), \qquad (2)$$

is a correction of the state-action value, *Q(s,a)*. The quantity *r* represents the reinforcement received for state transition from *s* to *s'*. The coefficient γ is called a discount factor and determines the value of future corrections. If it is 0, only immediate rewards are maximized. A value close to 1 implies more focus on long term reward maximization. Following the TD(0) method used here, the state-action values are updated by the following formula

$$Q^{k+1}(s,a) = Q^{k}(s,a) + \alpha\left(r + \gamma\, Q^{k}(s',a') - Q^{k}(s,a)\right). \qquad (3)$$

The learning rate, α, is selected from an interval *0* to *1* and monotonically decreased as the learning progresses.

After the state-action values update is complete, a new policy is extracted from the Q-table using the ε-greedy approach [12]. If we represent the policy by $\pi(s)$, the best rated actions are selected with probability *1-*ε, (for ε between 0 and 1),

$$\pi(s) = \arg\max_{a} Q(s,a). \qquad (4)$$

Conversely, random actions are selected with probability ε. The ε-greedy approach enables exploration of the state-action space, which is a very important component of the optimization process. To be able to use SARSA for the dose optimization in MPC-based drug delivery, we represent the minimization of the cost function as the dual problem of maximizing the long term reinforcement.

We formulate the learning problem as follows. The state vector *s* contains two components, current hemoglobin level, *Hgb*, and last rHuEPO dose, *EPO*. The recommended rHuEPO adjustment, Δ *EPO*, is the action. We use the following reinforcement

$$r = 1 - 4\,(11.5 - Hgb)^2. \qquad (5)$$

This function sends a positive reinforcement to any action driving the hemoglobin levels to, or maintaining it within the range *11* to *12* g/dL. If the hemoglobin level is outside this range, the corresponding action is penalized. This reinforcement promotes hemoglobin levels close to the median of the target range, *11.5* g/dL. The goal of the policy is to maximize the sum of the reinforcements over a period of time, H_p. The optimization episodes are repeated for a pre-specified number of times, decreasing the learning rate α, and the probability of selecting a random action, ε, after each episode.

III. Experimental Results

We collected data from *105* hemodialysis patients treated at the Division of Nephrology, University of Louisville in the year 2005. The data contained monthly hemoglobin levels and rHuEPO doses. We used these data to develop dose-response models of the patients, as described in [10]. We implemented the models as Multilayer Perceptron networks predicting the hemoglobin levels one month ahead, based on past *3* monthly rHuEPO doses and hemoglobin levels. We divided the data into *51* random, equally sized training and testing sets, and created a single model for each data set combination. We then randomly selected one of the models to serve as the MPC predictive model, and used the remaining *50* to simulate individual patients. This emulated the mismatch between the MPC model and the patient that occurs in real life. By using the same predictive model for all simulated patients, we mimicked the population approach to drug administration.

We set the length of a single optimization episode, H_p, to *12* months. The effect of a single rHuEPO dose can last as long as *2* months. We decided that *12* months was sufficient time length to thoroughly evaluate a single policy. We set the learning rate, α, to *0.9*, the discount rate, γ, to *0.99*, and the probability of selecting a random action, ε, to *99%*. The number of optimization episodes during one MPC step was determined through trial and error. We found *1200* episodes sufficient to achieve the convergence of the dosing policy. We decreased α by the factor of *0.995* and ε by *0.1%* after each episode. We limited the minimum value of ε to *1%*. We evaluated the MPC for each simulated patient over a period of *12* months, i.e. rHuEPO dose adjustment intervals.

To establish a benchmark, we simulated the anemia management using an algorithmic implementation of the clinical protocol (AMP) used at the Division of Nephrology in the year 2002, on the same *50* patient models described above.

The simulation results are illustrated in Figures 3 and 4 and summarized in Table I. We found that *10* out of the simulated *50* patients achieved hemoglobin levels above target range without receiving any rHuEPO. We decided not to include these individuals in the statistical analysis, as neither the protocol, nor the proposed MPC algorithm influenced their hemoglobin level. The top graph in Figure 3 shows a time plot of mean hemoglobin levels in the *40* patients (continuous line) and their standard deviations (whiskers) when the rHuEPO doses recommendations are produced by the MPC algorithms. The bottom graph shows the mean rHuEPO dose recommendations (continuous line) and their standard deviations. Figure 4 shows the same data for dose recommendations produced by the simulated AMP. These figures and Table I show that the proposed MPC approach and the AMP achieve the same mean hemoglobin

level in the simulated patient population. Compared to the AMP, the MPC achieves marginally better hemoglobin variability and rHuEPO utilization. Figure 3 also shows that the dose recommendations produced by the MPC are much more consistent than those produced by the AMP. The mean hemoglobin level achieved by the MPC shows a tendency to converge toward the target range. The mean hemoglobin level achieved by the simulated AMP fluctuates around the upper boundary of the target range. The inability to maintain the hemoglobin within the target range by the AMP can be attributed to the fact that we used a version from the year 2002. On the other hand, the patient models were created from data collected in 2005.

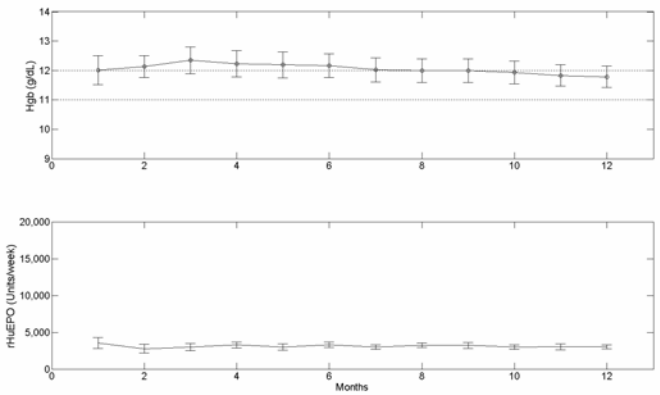

Fig. 3. Hemoglobin response within the simulated population (top) and rHuEPO dose as recommended by the MPC (bottom). Continuous lines represent mean values, whiskers represent standard deviations. Target hemoglobin range is shown by dotted lines.

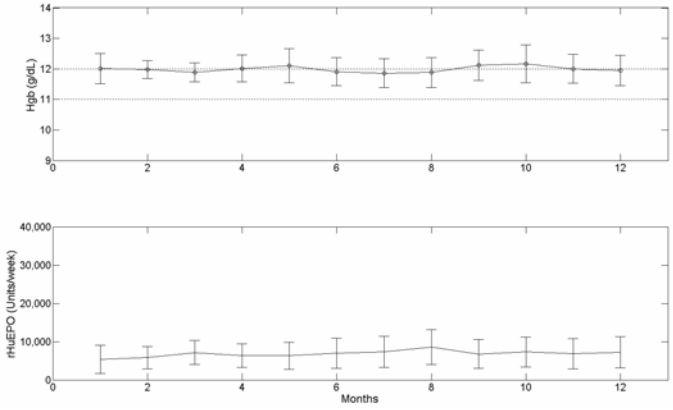

Fig. 4. Hemoglobin response within the simulated population (top) and rHuEPO dose as recommended by the AMP (bottom). Continuous lines represent mean values and the whiskers represent standard deviations. Target hemoglobin range is shown by dotted lines.

TABLE I
STATISTICAL SUMMARY OF THE SIMULATIONS (N = 40)

Method	AMP	MPC
Average Hgb (g/dL)	12.0 [10.1 13.80]	12.0 [10.4 13.70]
Hgb variability (g/dL)	0.52 [0.0 1.28]	0.49 [0.0 0.98]
Total rHuEPO (Units)	328,000	299,000

IV. CONCLUSIONS

We proposed an approach to computer-assisted drug delivery combining Model Predictive Control for simulation of patient response and Reinforcement Learning for optimization of the dosing strategy. We evaluated this approach through numerical simulations of anemia treatment with recombinant human erythropoietin using patient models created from clinical data. We show that the proposed algorithm performs as well as the clinical protocol for anemia management in terms of mean hemoglobin level and improves upon the protocol in terms of hemoglobin variability.

REFERENCES

[1] S. Vozeh and L. Steimer, "Feedback Control Methods for Drug Dosage Optimization: Concepts, Classification and Clinical Application," *Clinical Pharmacokinetics*, vol. 10, 1985, pp. 457-476.

[2] Z. Trajanoski and P. Wach, "Neural Predictive Controller for Insulin Delivery Using the Subcutaneous Route," *IEEE Trans. Biomed. Engineering*, Vol. 45, No. 9, September 1998, pp. 1122-1134.

[3] J.Y. Conway and M.M. Polycarpou, "Using Adaptive Networks to Model and Control Drug Delivery," *IEEE Intelligent Control*, April 1996, pp. 31-37.

[4] M.M. Polycarpou and J.Y. Conway, "Indirect Adaptive Control of Drug Delivery Systems," *IEEE Trans. Automatic Control,* Vol. 43, No. 6, June 1998, pp. 849-856.

[5] A.E. Gaweda, A.A. Jacobs, G.R. Aronoff, M.E. Brier, "Intelligent control for drug delivery in management of renal anemia," in *Proc. 2004 Int. Conf. Machine Learning and Applications*, pp. 355-359.

[6] A.E. Gaweda, M.K. Muezzinoglu, G.R. Aronoff, A.A. Jacobs, J.M. Zurada, M.E. Brier, "Reinforcement Learning Approach to Chronic Pharmacotherapy," in *Proc. 2005 IEEE Int. Joint Conf. Neural Networks,* vol. 5, pp. 3290-3295.

[7] A.E. Gaweda, M.K. Muezzinoglu, G.R. Aronoff, A.A. Jacobs, J.M. Zurada, M.E. Brier, "Individualization of Pharmacological Anemia Management using Reinforcement Learning," *Neural Networks*, vol 18, 2005, pp. 826-834.

[8] J.W. Eschbach and J.W. Adamson, "Anemia of end-stage renal disease (ESRD)," *Kidney Int.,* 28, 1985, pp. 1-5.

[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] 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 Networks*, Vol. 16, 5-6, 2003, pp. 841-845.

[11] D.W. Clarke, C. Mohtadi, and P.S. Tuffs, "Generalized predictive control--I. The basic algorithm," *Automatica*, 23, 1987, pp. 137--148.

[12] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction,* MIT Press: Cambridge, MA, 1998.

[13] R.R. Negenborn, B. De Schutter, M.A. Wiering, and H. Hellendoorn, "Learning-based model predictive control for Markov decision processes," *Proc. 16th IFAC World Congress*, Paper 2106 / We-M16-TO/2.

[14] C. Watkins, *Learning from Delayed Rewards*, Thesis, University of Cambidge, England, 1989.