Optimal Medication Dosing from Suboptimal Clinical Examples: A Deep Reinforcement Learning Approach

Shamim Nemati[1,†,*], Mohammad M. Ghassemi[2,†], and Gari D. Clifford[1,3]

Abstract— **Misdosing medications with sensitive therapeutic windows, such as heparin, can place patients at unnecessary risk, increase length of hospital stay, and lead to wasted hospital resources. In this work, we present a clinician-in-the-loop sequential decision making framework, which provides an individualized dosing policy adapted to each patient's evolving clinical phenotype. We employed retrospective data from the publicly available MIMIC II intensive care unit database, and developed a deep reinforcement learning algorithm that learns an optimal heparin dosing policy from sample dosing trails and their associated outcomes in large electronic medical records. Using separate training and testing datasets, our model was observed to be effective in proposing heparin doses that resulted in better expected outcomes than the clinical guidelines. Our results demonstrate that a sequential modeling approach, learned from retrospective data, could potentially be used at the bedside to derive individualized patient dosing policies.**

I. Introduction

Deviations from established treatment protocols in complex clinical environments, such as the intensive care unit (ICU), are a common and necessary component of effective treatment. While some of these deviations are errors [1], many more are innovative adjustments made by clinicians to adapt treatments to the individualized (perceived) needs of patients. Clinicians often reference clinical context, patient preference, provider bias, prior training, local medical practice, and lack of (or conflicting) randomized clinical trials (RCTs)-based evidence as the driving factors of the variability in treatment approach.

Medication dosing is one example of treatment policy where deviations from the norm are common, and sometimes useful. However, som medications have highly sensitive therapeutic windows, making them easily over- or under-dosed. Mismanagement of such drugs can (1) drive up costs by unnecessarily extending hospital length of stay (2) reduce hospital productivity by requiring additional follow-up interventions to correct for mistakes and, (3) in some cases, place already frail patients at risk of additional complications [2], [3].

Unfractionated Heparin (UH), is one example of such a medication - leading to increased risk of bleeding if overdosed, and increased risk of clot formation if underdosed [4]. The same risks exist in the case of Warfarin, (a drug used by approximately 20 million patients in the United States alone) which is estimated to be incorrectly dosed in a staggering one-third of patients [5]. The sensitivities of patients to these drugs, and many other drugs with misdosing consequences, demands the development of robust decision support tools, which will consider a greater breadth and depth of factors that influence patient outcomes. Importantly, such tools should be designed to minimize the number of treatment iterations required to bring patients to the therapeutic state as defined by the caretaker, not to dictate those definitions themselves.

As retrospective clinical archives continue to grow in both breadth (e.g. through multi-center initiatives) and depth (using higher resolution data), access to variations in patient characteristics, and corresponding treatment decisions, has provided an unprecedented opportunity to generate tools that may learn optimal personalized treatment policies from the data. In this work, we present a *reinforcement learning* (RL) algorithm for learning actionable policies to minimize dosing errors for dosing medications with sensitive therapeutic windows. For this article we focus on the dosing of UH as a useful illustrative example.

RL is particularly well-suited for the medication dosing problem given the sequential nature of clinical treatment - where multiple treatment decision are performed without immediate knowledge of effectiveness. Indeed, the lack of a one-to-one correspondence between actions and outcomes makes it difficult to assign credit or blame to individual actions along the way to an intermediate or terminal outcome. Moreover, the effect of interventions for a given patient can be non-deterministic, and attempting to predict the effects of a series of treatments over time only raises to this uncertainty.

Within the RL literature this type of problem is known as a *credit assignment task* and may be modeled using a Markov decision processes (MDP) for probabilistic inference over time, given non-deterministic action effects. In our case specifically, we use a partially observable Markov decision process (POMDP), which extends a standard MDP by modeling an internal belief about patient state and their expected response to interventions. In this work, we provide approximate dosing solutions in both discrete action spaces [6].

II. Methods

In many clinical settings, UH dosing begins with the intravenous administration of a weight-based dosage of heparin [3]. After 4 to 6 hours, a laboratory test of blood clotting is performed to determine the activated partial thrombo-

[1]Dept. of Biomedical Informatics, Emory University, Atlanta, GA 30322.
[2]Dept. of Electrical Engineering and Computer Science, MIT, Cambrige, MA 02139.
[3]Dept. of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.
† Equal contribution authors
*Corresponding author, Email: shamim.nemati@emory.edu

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

plastin time (aPTT). This test serves as feedback (or a *reward/reinforcer* in the RL literature) for the clinicians. Given the feedback, a decision is made to increase, decrease or maintain the heparin dosage (see Fig. 2, panel (a)) until the next aPTT measure.

Our goal is to infer an optimal dosing strategy that accounts for both the aPTT level, and evolving patient physiological condition. To accomplish this inference, we train a RL model using the time series of several common clinical measurements within the patient's electronic medical record (EMR).

The RL model utilizes a cohort of N patients, with associated multivariate time series of laboratory values Y, clinical actions a, and associated rewards r: $\{(Y^{(1)}_{M\times T_1}, a^{(1)}_{1\times T_1}, r^{(1)}_{1\times T_1}), \cdots, (Y^{(N)}_{M\times T_n}, a^{(N)}_{1\times T_1}, r^{(N)}_{1\times T_1})\}$, where the n-th time series $Y^{(n)}$ may be of length T_n and include M channels, $a^{(n)}$ can be a discrete (categorical) or continuous time series of actions of length T_n, and $r^{(n)}$ is a one dimensional time series of rewards of length T_n, with up to $T_n - 1$ missing values (delayed rewards).

Within the RL framework [7], a reward is a measure of the *immediate* utility of an action in a given state (a summary of all we can know about the agent's environment). The RL agent's objective is to maximize its expected long-term reward by following a policy $\pi : S \rightarrow A$, where S denotes the state-space, and A denotes the set of possible actions. Thus, our goal is to simultaneously learn the state sequence $s^{(n)}$ associated with each time-series $Y^{(n)}$, and an optimal policy $\pi^*(s^{(n)}_t)$ that suggests an action $a^{(n)}_t$ with maximum expected long-term reward. We define a *clinician-in-the-loop* policy, as a control policy that takes into account the action of a clinician as well as her patient's response when suggesting a new action. In this setting, a clinician may choose to approve or overwrite the action suggested by the RL agent at any given point in time.

A. Dataset

Following Ghassemi et al. [8], we extracted data for 4470 patients from the publicly available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II database [9] that received a heparin intravenous infusion at some point during their ICU stay (randomly assigned to 80% training and 20% testing sets).

We extracted up to 48 hours of data per patient, starting from the time of first heparin administration. Our extracted features included a comprehensive laboratory measurements: heparin dose level and aPTT measurements calculated over the four hours prior to the selected time ($t-4:t$, in hours), arterial carbon dioxide level (CO_2), heart rate (HR), heparin dose ($t-4:t$, in hours), albumin, systolic and diastolic arterial blood pressure (SBP and DBP), bilirubin, creatinine, Glasgow Coma Score (GCS), hematocrit, hemoglobin, International normalized ratio of prothrombin (INR), blood PH, platelet count, prothrombin time, respiration rate, oxygen saturation of arterial blood (SA02), daily Sequential Organ Failure Assessment (SOFA) scores, temperature, troponin, urea, and white blood cell count (WBC). Additionally, we collected the following dichotomous features: ethnicity (white/non-white), ICU service type (surgical/medical), gender, transfer from another hospital, pulmonary embolism and obesity. We also extracted patient age and weight.

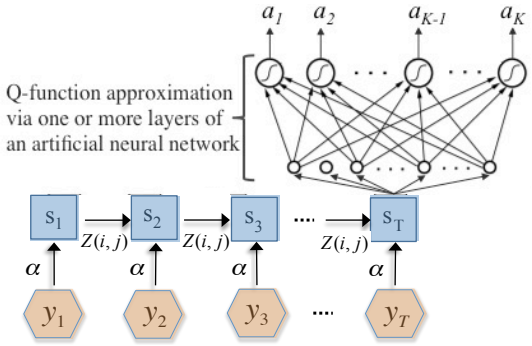

Fig. 1. A discriminative HMM (DHMM) with an added Q-learning layer consisting of a two-layer neural network (aka, a Q-Network [6]). For clarity, we only show the Q-network on the last node, but otherwise at each point in time the Q-network estimates the long-term value of each of the K possible actions given the marginal state probability vectors.

The therapeutic range of anticoagulation was defined as an aPTT between 60 and 100 seconds [8].

B. Exclusion Criteria and Pre-Processing

We excluded all patients which were transferred from another institution. All feature data were re-sampled at an hourly rate (where multiple measurements within the same hour window were replaced by their median value). To account for missing hourly values, we utilized sample-and-hold interpolation which we consider the most practical form of interpolation at the bedside, given the non-random (and generally unknown) nature of the missing data in most clinical settings.

C. Reinforcement Learning

The objective of the RL medication dosing agent is to learn a dosing policy that maximizes the overall fraction of time a given patient stays within his/her therapeutic aPTT range. We constructed a RL reward function reflecting this objective: $r_t = \frac{2}{1+e^{-(aPTT_t-60)}} - \frac{2}{1+e^{-(aPTT_t-100)}} - 1$. This function assigns a maximal reward of one when a patients aPTT value is within the therapeutic window which rapidly diminishes towards a minimal reward of -1 as the distance from the therapeutic window increases.

Since the actual physiological state of the patient is at best only partially observed, the agent has to infer both the state of the patient and an optimal policy from sample trajectories of its interaction with the environment, i.e., the recorded data within the EMR.

1) Discriminative Learning for State Estimation: In the most general case, one may use a dynamic Bayesian network (DBN) or a recurrent neural network (RNN) to infer a sequence of hidden states associated with the observed time series of clinical parameters. For simplicity, here we utilize a discriminative hidden Markov model (DHMM) for state estimation (as shown in Fig. 1). Since we are interested in inferring states that can assist in learning a policy for

maximizing the long-term reward, we employ gradient-based supervised learning [10] to learn the parameters of the DHMM (collectively called $\boldsymbol{\beta}$) so to directly minimize the RL cost function, as described next.

2) Discrete Action Spaces: Watkins's *Q-learning* algorithm [11] works by learning an action-value function that provides the expected long-term reward of taking a given action in each state. The *Q-function* for a state-action pair is defined as [7]:

$$Q(s_t^{(n)}, a_t^{(n)}) = \max_\pi \mathbf{E}[r_t^{(n)} + \gamma r_{t+1}^{(n)} + \gamma^2 r_{t+2}^{(n)} + \cdots],$$

where γ ($\in [0,\ 1]$) is a discount factor, and the maximum is taken over all possible policies π. Within the *fitted Q-learning* framework [12] the Q-function is represented by a neural network with weights W. A parametrized version of Eq. (1) using dynamic-programming can be written as [7]:

$$Q^*(s_t^{(n)}, a_t^{(n)}; W) = \mathbf{E}_{s'}[r_t^{(n)} + \gamma \max_{a' \in A} Q^*(s', a'; W)],$$

where $s_t^{(n)}$ is a shorthand for the marginal $P_{\boldsymbol{\beta}}(s_t^{(n)}|y_{1:t}^{(n)})$, and the subscript $\boldsymbol{\beta}$ is used to make the dependence of the state estimate on the parameters of the DHMM explicit. The first term within the expectation operator in the above equation is the immediate reward of taking action $a_t^{(n)}$ in state $s_t^{(n)}$, and the second term is the discounted long-term reward the agent can expect by taking the best action thereafter. Given the optimal Q-function, the optimal policy is given by $\pi^*(s_t^{(n)}; W) = \text{argmax}_{a' \in A} Q^*(s_t^{(n)}, a'; W)$. The Q-learning algorithm updates the weights W by minimizing the following cost function [6]:

$$L(W_{i+1}) = \frac{1}{2|N_i|} \sum_{n \in N_i} \sum_{t=1}^{T^{(n)}} [v_t^{(n)}(W_i) - Q(s_t^{(n)}, a_t^{(n)}; W_{i+1})]^2,$$

where $v_t^{(n)}(W_i) = r_t^{(n)} + \gamma \max_{a' \in A} Q(s_{t+1}^{(n)}, a'; W_i)$ is the expected value of the state-action pair under the current Q-function at time t and for the example n within the current training batch ($N_i \subseteq \{1, \cdots, N\}$). Note that, we have replaced $Q^*(.,.;W)$ by its best current estimate $Q(.,.;W)$; this is a form of bootstrapping. The gradient of this cost function with respect to the weights W_{i+1} is given by:

$$\frac{\partial L(W_{i+1})}{\partial W_{i+1}} = -\sum_{n \in N_i} \sum_{t=1}^{T^{(n)}} [v_t^{(n)}(W_i) - Q(s_t^{(n)}, a_t^{(n)}; W_{i+1})] \times \frac{1}{|N_i|} \frac{\partial Q(s_t^{(n)}, a_t^{(n)}; W_{i+1})}{\partial W_{i+1}}.$$

3) Optimization: The gradient of the RL cost function with respect to the network weights (W) can be further backpropagated through the DHMM parameters ($\boldsymbol{\beta}$) (or any DBN or RNN model used for state estimation) [10]. The resulting combined gradients can be directly plugged into an optimization package, such as MATLAB's *minFunc* [13], to optimize all model parameters simultaneously (i.e., end-to-end supervised training). When optimizing over a large patient cohort, we found that a stochastic optimization approach—using mini-batches with a few iterations per batch and a momentum term—yielded improved generalization performance with significant speed up. Hyperparameters of the DHMM and the neural network representing the policy (such the number of layers and nodes) were tuned using Bayesian Optimization [14].

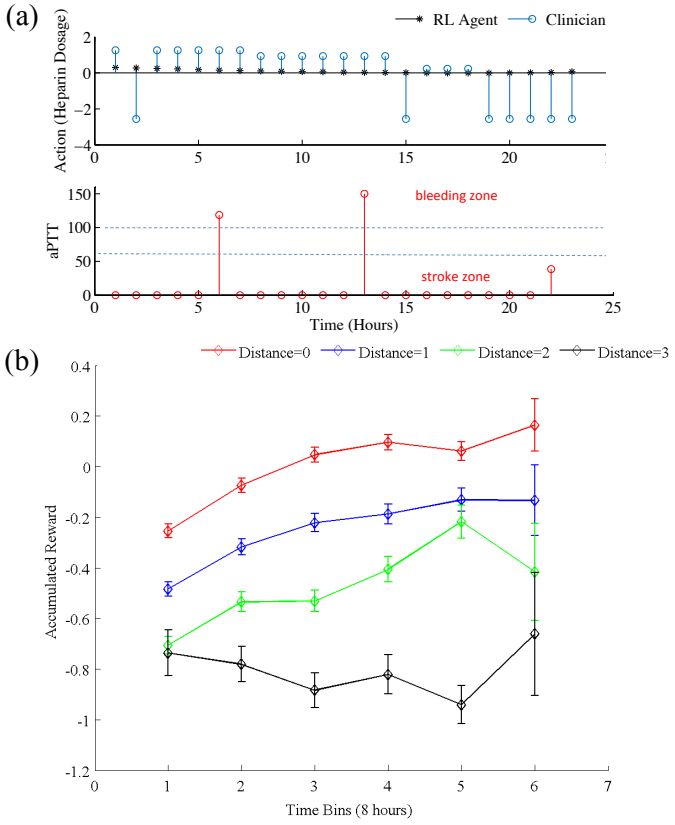

Fig. 2. Panel (a): each heparin dosing trial starts with an initial dosing of heparin (stemmed open circles) followed by sequential adjustments over the next 24-48 hours upon availability of new aPTT test results (middle plot) and according to a hospital dosing protocol/policy (π_H). The prescribed actions by the trained RL agent (π_{RL}) are superimposed (stemmed asterisks). Panel (b): Quantifying dosing performance (accumulated reward; mean and standard errors) over 8 hour time bins ($\times$ 6 bins = 48 hours), and distance from RL policy (color-coded): $|\pi_H - \pi_{RL}|$. These results shows that, consistently over time, adherence to the RL policy (Red lines; distance of zero) results in the highest accumulated reward.

III. Results, Discussion and Future Work

We discretized the heparin values using six quantile intervals to define a discrete set of actions. Fig. 2 (a) shows an example of heparin dosing (mean-normalized) by a clinician, and the corresponding recommended dosing of the RL agent. Note that the clinician initially over-doses the patient, as reflected in the aPTT measurements 6 hours through the trial. The situation worsens till 15 hours through the trial when a corrective action is finally made; that is, taking the patient completely off the heparin for one hour, followed by three hours of consecutive heparin administration at the population mean level (zero level), and finally patient is completely taken off of heparin over the next five hours (19-23 hours). However, the corrective action results in an under-dosing of the patient, as reflected by the last aPTT measurement

(22 hours through the trial). In comparison, the trained RL agent's recommendation starts slightly above the population mean for heparin and then converges to the population mean, which is likely to bring patients within their therapeutic range more quickly.

To further test this hypothesis, we grouped each instance of heparin administration according to its distance from the dosage recommended by our trained RL agent. Thus, a distance of zero indicates that the clinically administered dose matched the RL agent's recommendation. The testing set results presented in Fig. 2(b) shows that, on average and consistently over time, following the recommendations of the RL agent (red line) results in the best long-term performance. In fact, while the expected reward over all dosing trajectories in our cohort is negative, patients whose administered heparin trajectory most closely followed the RL agent's policy could on average expect a positive reward after just a few adjustment.

To our knowledge this is the first successful application of deep reinforcement learning to the problem of medication dosing in the ICU. Application of RL to design of clinical trials and adjustment of clinical treatments have been previously suggested in the literature [15], [16], [17]. However, the previous works do not combine state estimation and RL training via end-to-end optimization. In our experience, feeding the raw laboratory values to the RL algorithm did not perform well. This is likely due to the high-dimensional nature of clinical observations often made at the bedside. State estimation provides a summary of these measurements in terms of a few compact state variables that are continuously updated as more measurements become available.

In spite of the great success of deep reinforcement learning in other fields [6], application of this technique to clinical problems has been limited, since large clinical datasets with granular temporal data are often scarce. With the advent of big clinical datasets and more efficient algorithms, including state estimation, better initialization, and optimization, we may start to tackle the problems we face in this domain. The results presented in this work illustrates that a data-driven approach to heparin dosing performs better, on average, than the state-of-the-art in clinical practice.

The example presented here should be taken as illustrative. Whether the suboptimal heparin dosing we observed were from intentional actions on the part of the clinician, mistakes, or simply due to a lack of adherence to hospital guidelines are beyond our ability to investigate with the dataset at hand. This points at one of the major challenges of retrospective analysis of clinical big data; the rational for treatment decisions are often unknown, and some features which may be important for understanding outcomes may be missing, most likely not at random. Nevertheless, one major advantage of retrospective analysis is the low cost, high volume, and scalability. More importantly, retrospective data often provides diverse representations of the critically ill, including members of the population which might be too ill to include in a clinical trials. Hence, there are some areas of research that, in the interest of ethics, can only be carried out retrospectively. Nevertheless, the application of machine learning and in particular sequential decision making techniques to medicine is still at its infancy, and we believe advances in deep reinforcement learning will play an important role in the future of precision medicine and achieving a learning health care system [18].

ACKNOWLEDGMENTS

S.N. is grateful for an NIH early career development award in biomedical big data science (1K01ES025445-01A1). M.G. Is grateful for the support of the Salerno foundation, which supported his PhD studies. The authors would like to thank Professor Ryan Adams (Harvard-SEAS) and Dr. Roger Mark (MIT-EECS) for their insightful comments.

REFERENCES

[1] J. T. James, "A new, evidence-based estimate of patient harms associated with hospital care," *Journal of patient safety*, vol. 9, no. 3, pp. 122–128, 2013.

[2] S. Alban, "Adverse effects of heparin," in *Heparin-A Century of Progress*. Springer, 2012, pp. 211–263.

[3] R. A. Raschke, B. Gollihare, and J. C. Peirce, "The effectiveness of implementing the weight-based heparin nomogram as a practice guideline," *Archives of internal medicine*, vol. 156, no. 15, p. 1645, 1996.

[4] C. S. Landefeld, E. F. Cook, M. Flatley, M. Weisberg, and L. Goldman, "Identification and preliminary validation of predictors of major bleeding in hospitalized patients starting anticoagulant therapy," *The American journal of medicine*, vol. 82, no. 4, pp. 703–713, 1987.

[5] H. Q. Ontario *et al.*, "Point-of-care international normalized ratio (inr) monitoring devices for patients on long-term oral anticoagulation therapy: an evidence-based analysis," *Ontario health technology assessment series*, vol. 9, no. 12, p. 1, 2009.

[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[8] M. M. Ghassemi, S. E. Richter, I. M. Eche, T. W. Chen, J. Danziger, and L. A. Celi, "A data-driven approach to optimized medication dosing: a focus on heparin," *Intensive care medicine*, vol. 40, no. 9, pp. 1332–1339, 2014.

[9] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): A public-access intensive care unit database," *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.

[10] S. Nemati, R. Adams *et al.*, "Supervised learning in dynamic bayesian networks," in *NIPS Workshop on Deep Learning and Representation Learning*, 2014.

[11] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[12] M. Riedmiller, "Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method," in *Machine Learning: ECML 2005*. Springer, 2005, pp. 317–328.

[13] M. Schmidt, "minfunc: unconstrained differentiable multivariate optimization in matlab," 2012.

[14] M. Ghassemi, L.-w. Lehman, J. Snoek, and S. Nemati, "Global optimization approaches for parameter tuning in biomedical signal processing: A focus on multi-scale entropy," in *Computing in Cardiology Conference (CinC), 2014*. IEEE, 2014, pp. 993–996.

[15] B. Chakraborty and S. A. Murphy, "Dynamic treatment regimes," *Annual review of statistics and its application*, vol. 1, p. 447, 2014.

[16] H. Asoh, M. S. S. Akaho, T. Kamishima, K. Hasida, E. Aramaki, and T. Kohro, "An application of inverse reinforcement learning to medical records of diabetes treatment," 2013.

[17] M. K. Bothe, L. Dickens, K. Reichel, A. Tellmann, B. Ellger, M. Westphal, and A. A. Faisal, "The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas," *Expert review of medical devices*, vol. 10, no. 5, pp. 661–673, 2013.

[18] B. D. Friedman CP, Wong AK, "Achieving a nationwide learning health system," *Sci Transl Med*, vol. 2, no. 57, pp. 1–3, 2010.