Reinforcement Learning for Pharmacometrics: A Proof of Concept and Future Directions

Annaliese Wieler, Ph.D., Matthew Wiens, M.A., Samuel P. Callisto, Ph.D., Megan Cala Pane, Ph.D., Hillary Husband, Ph.D.
Metrum Research Group, Boston, MA, USA

Background

Reinforcement Learning (RL)

- An **agent** learns a **policy** to optimize a **reward** given that it is in a certain state.
  - The **agent** is the machine learning model.
  - The **policy** is a decision that changes the current state of the system.
  - The **reward** is a function that gives positive value for desired actions and negative values for undesired actions.
- RL is advantageous over traditional supervised machine learning situations where the outcome is unclear, especially when the outcome is probabilistic due to random sources of variation.
- RL is commonly used to create game players (i.e., AlphaGo) which require averaging over a large number of potential outcomes for a large number of potential policies to determine the correct one.
- RL has recently begun to be used within pharmacometrics for precision dosing in quantitative systems pharmacology (QSP) and real-time dosing, and there is a wide array of potential applications of RL within the field of pharmacology [1].

Challenges

- RL is extremely sensitive to the design of the reward function, and there is no standard methodology for defining the reward function.
- Integration of pharmacodynamic (PD) outcomes with pharmacokinetic (PK) profiles is ideal for defining the reward, but this adds an additional layer of uncertainty.
- RL requires very large training data sets, which is often a challenge in PK-PD analyses.

Opportunities

1. RL can be used to average over unknown sources of variation (such as inter-individual variability (IIV)).
2. RL can be used for sequential decision making such as dose adjustment based on observed outcomes and future expected rewards.

Objectives

Perform a proof-of-concept study to apply RL to a PK problem and identify future areas of research.

- Design a reward function that is capable of balancing penalization of both subtherapeutic and supratherapeutic exposures (minimum concentration in the dose interval (C_{trough}) and area under the concentration-time curve (AUC)).
- Identify future use cases of RL for dose optimization.

Methods

- A workflow was created to train a reinforcement learner to provide an optimal dose regimen based on individual sets of covariates (Figure 1).
  - The optimal vancomycin dose for the treatment of methicillin-resistant *Staphylococcus aureus* (MRSA) is recommended to lead to PK profiles with a C_{trough} between 15 and 18.2 mg/L and an AUC between 400 and 600 mg*hr/L [2].

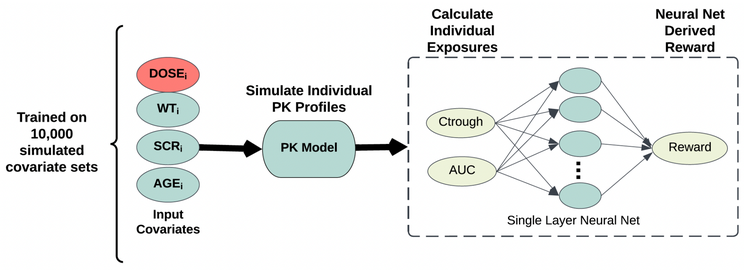

Figure 1. Process schematic of neural net (NN) reinforcement learner. Covariate sets included individual weight (WT_i), serum creatinine (SCR_i), age (AGE_i), and dose ($DOSE_i$).

- A PK model with covariate effects and IIV was used to simulate PK profiles for selected doses of vancomycin [3].
  - A two-compartment model for vancomycin was used to generate the exposure metrics used in the reinforcement learner.
  - A weight effect was included on both central and peripheral volume of distribution. The effects of maturation, age, and serum creatinine were included on clearance.
- A training data set of 10,000 individuals composed of model covariates (age, weight, and serum creatinine) was generated from NHANES [4]. Doses between 20-35 mg/kg were randomly assigned to each simulated subject, and the resulting PK profiles and associated reward were calculated.
- A neural net was fit in R using TensorFlow [5] with the model covariates and dose as predictors and the reward as the outcome.
- Separate reward functions for C_{trough} and high AUC were designed to reward exposures in the optimal range and penalize exposures known to lead to toxicities based on literature (Figure 2) [2].

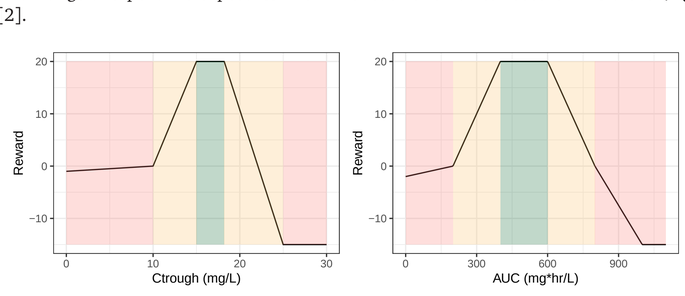

Figure 2. Final reward function used in RL. The reward function selected used a combination of linear and step functions. The green shaded regions indicate ideal exposure ranges, while yellow and red shaded regions indicate sub-optimal and dangerous exposure ranges.

Methods (continued)

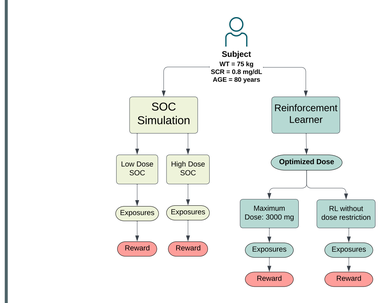

Figure 3. Simulation workflow for an example subject.

- 50 replicates for each of 100 new virtual subjects were simulated with covariate sets drawn from the NHANES dataset [4].
- PK profiles and the resulting rewards were generated both by the reinforcement learner and under standard-of-care (SOC) regimens (Figure 3) . Both low-dose and high-dose standard-of-care (SOC) regimens were simulated [2].
  - Low-dose SOC was simulated as a single 20 mg/kg dose.
  - High-dose SOC was simulated as a single 35 mg/kg dose.
- Both SOC regimens were based on exact body weight, rounded to the nearest 250 mg increment, and was not allowed exceed 3000 mg.
- Dose recommendations were taken directly from the reinforcement learner. An additional scenario was tested which capped dose recommendations from the reinforcement learner to a maximum dose of 3000 mg and discretized to 250 mg increments that constrained regimens within clinically implementable ranges.

Results

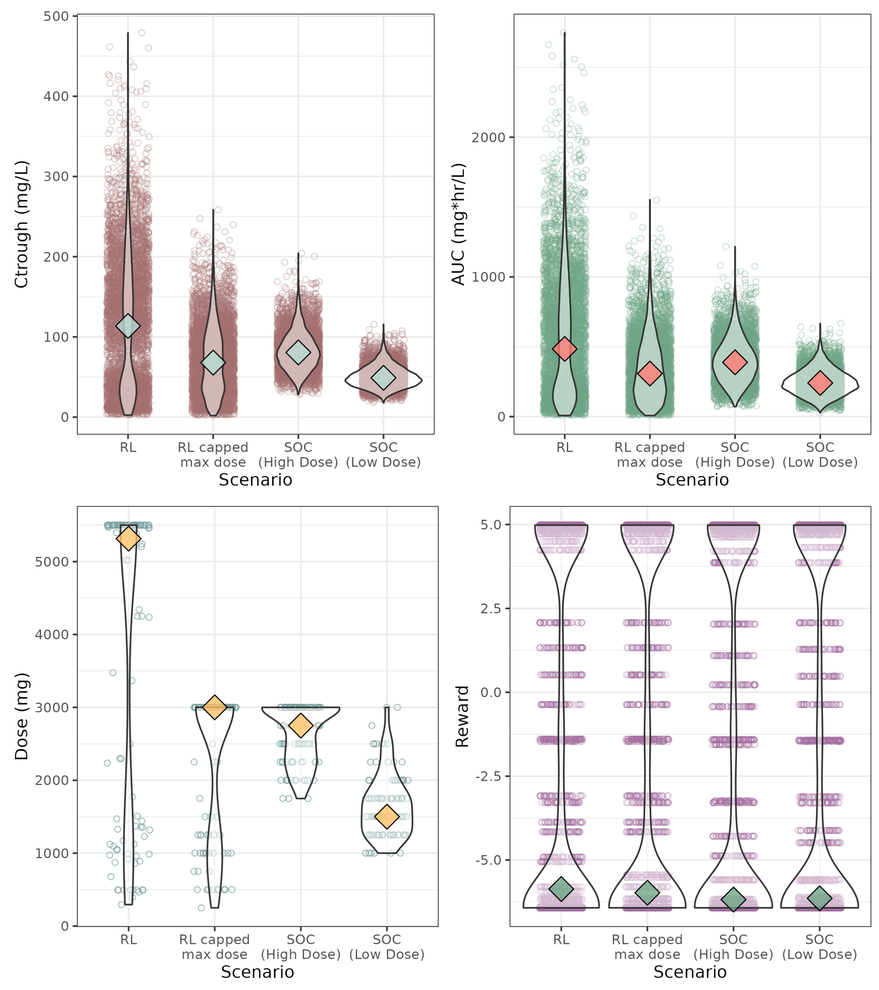

Figure 4. Exposure metrics, dose recommendation, and reward distribution for each scenario. The two relevant exposure metrics for the reward function were C_{trough} (mg/L) and AUC (mg*hr/L). Individual values overlaid as points over violin plot illustrating distribution of values. Within-scenario median included as overlaid triangle.

- Doses identified using the RL approach on average generated a larger reward (i.e., exposures within optimal range) when compared to SOC dose regimens (Figure 4, Table 1).
- Even when doses were capped to a maximum of 3000 mg, the RL dose regimens outperformed the SOC regimens (Figure 4, Table 1).
- The model was extremely sensitive to the choice of reward function.
  - The chosen reward function tended toward a higher-risk solution, pushing higher doses to maximize time in the ideal exposure range (Figure 4).
  - Alternative functions tested led to different outcomes, such as minimizing doses to reduce potential adverse effects.
- Doses proposed by the RL were generally different than those recommended by the weight-based SOC dose regimen (Figure 5), leading to different exposures over time (Figure 6).

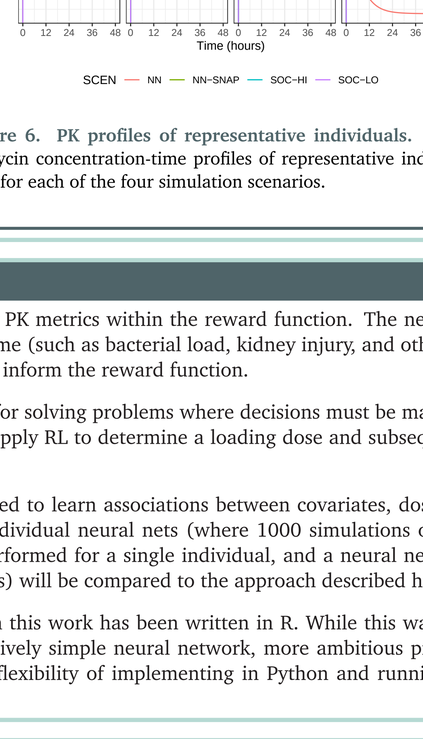

Figure 5. Comparison of doses proposed by each scenario.

		Median (Range)		
Scenario	Median reward	Dose (mg)	Ctrough (mg/L)	AUC (mg*hr/L)
NN	-5.87	5310 (294-5500)	114 (2.5-479)	484 (7.65-2750)
NN-SNAP	-5.97	3000 (250-3000)	67.9 (1.94-258)	308 (7.48-1550)
SOC-HI	-6.18	2750 (1750-3000)	80.6 (28.2-204)	389 (71.1-1220)
SOC-LO	-6.14	1500 (1000-3000)	48.9 (17.9-116)	241 (28-666)

Abbreviations: HI: high-dose regimen; LO: low-dose regimen; NN: neural network; SNAP: constrained maximum dose scenario; SOC: standard-of-care
Source code: standard-of-care-sims.R

Table 1. Summary of reward, dose, C_{trough}, and AUC by scenario.

Figure 6. PK profiles of representative individuals. Vancomycin concentration-time profiles of representative individuals for each of the four simulation scenarios.

Conclusions

- Reward functions that gave high rewards in the optimal range of C_{trough} and AUC and linearly decreasing rewards outside of the optimal range led to dose suggestions in line with clinical guidance.
- While allowing higher doses than the clinically-advised maximum led to higher reward values in some individuals, the rewards for RL-recommended optimal doses were still higher than those according to the SOC when the RL-recommended dose was capped at the maximal clinical dose.
- While the RL adopted a "higher risk, higher reward" strategy, the behavior could be altered by changing the form of the reward function.
- Although only the starting dose was identified by the approach shown here, this proof of concept motivates using this approach in selecting optimal maintenance dose regimens as well.

Future Directions

- This approach used only PK metrics within the reward function. The next iteration will consider a PD outcome (such as bacterial load, kidney injury, and other markers of safety and efficacy) to inform the reward function.
- RL has the most benefit for solving problems where decisions must be made sequentially. Future work will apply RL to determine a loading dose and subsequent maintenance dosing.
- A single deep RL was used to learn associations between covariates, doses, and rewards. In the future, individual neural nets (where 1000 simulations of dose and reward outcomes are performed for a single individual, and a neural net is built to model these relationships) will be compared to the approach described here.
- All code used to perform this work has been written in R. While this was sufficient to run a single and relatively simple neural network, more ambitious projects may benefit from the added flexibility of implementing in Python and running on GPU processors.

References

[1] Ribba, B., Dudal, S., Thierry, L. and Peck, R.W. Model-Informed Artificial Intelligence: Reinforcement Learning for Precision Dosing. *Clin. Pharmacol. Ther.* **107** (2020):853–857.

[2] Rybak, M.J., Le, J., Lodise, T.P., Levine, D.P., Bradley, J.S., Liu, C., Mueller, B.A., Pai, M.P., Wong-Beringer, A., Rotschafer, J.C., Rodvold, K.A., Maples, H.D. and Lomaestro, B.M. Therapeutic monitoring of vancomycin for serious methicillin-resistant Staphylococcus aureus infections: A revised consensus guideline and review by the American Society of Health-System Pharmacists, the Infectious Diseases Society of America, the Pediatric Infectious Diseases Society, and the Society of Infectious Diseases Pharmacists. *Am J Health Syst Pharm.* **77** (2020):835–864.

[3] Colin, P.J., Allegaert, K., Thomson, A.H., Touw, D.J., Dolton, M., de Hoog, M., Roberts, J.A., Adane, E.D., Yamamoto, M., Santos-Buelga, D., Martín-Suarez, A., Simon, N., Taccone, F.S., Lo, Y.L., Barcia, E., Struys, M.M.R.F. and Eleveld, D.J. Vancomycin pharmacokinetics throughout life: Results from a pooled population analysis and evaluation of current dosing recommendations. *Clin. Pharmacokinet.* **58** (2019):767–780.

[4] NHANES - National Health and Nutrition Examination Survey Homepage. `http://www.cdc.gov/nchs/nhanes/` (2018). Accessed: 2024.

[5] Abadi, M, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015). Software available from tensorflow.org.

Metrum Research Group Publications and Posters