UNIVERSITÉ TOULOUSE III PAUL SABATIER

INSTITUT de MATHEMATIQUES de TOULOUSE

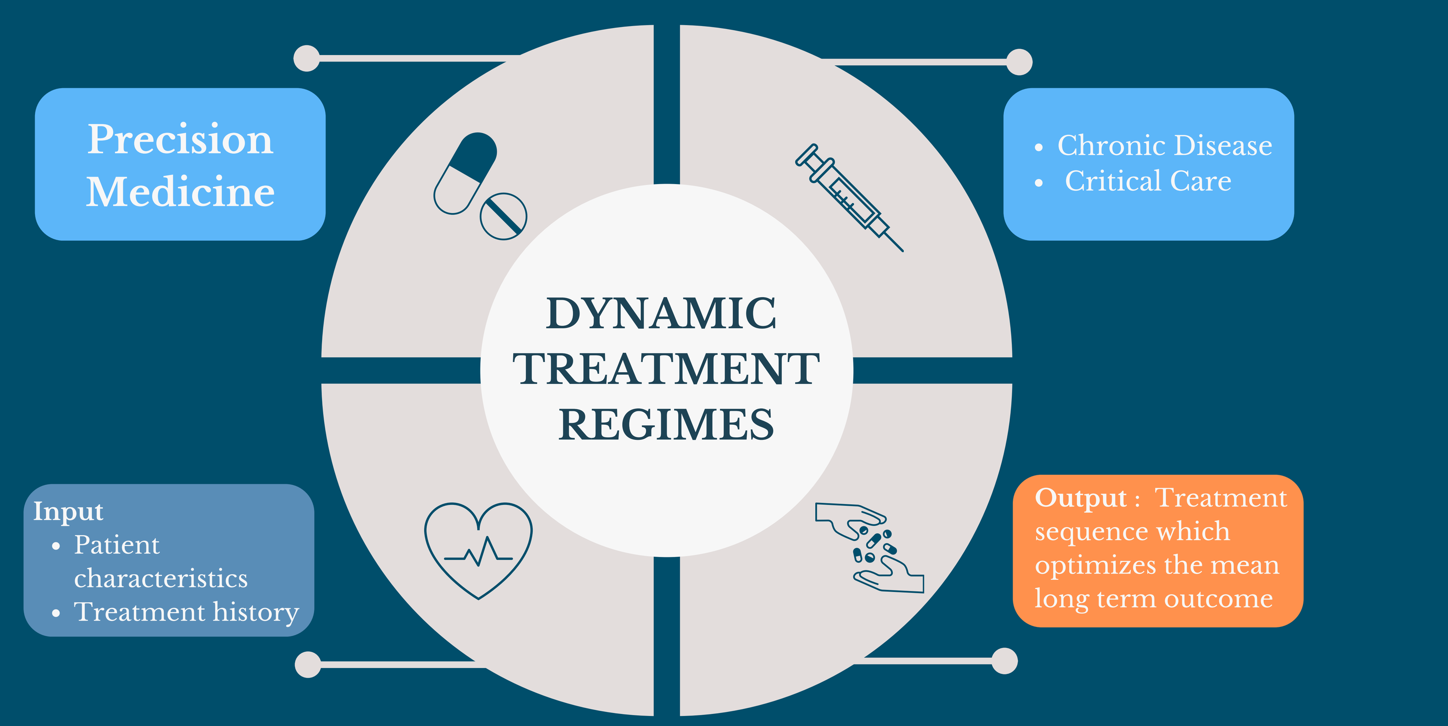
UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Integration of Medical Knowledge into Reinforcement Learning for Dynamic Treatment Regimes

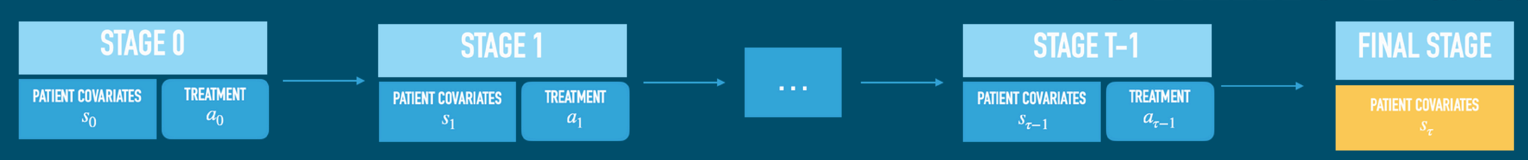
Sophia YAZZOURH

Nicolas SAVY, Philippe SAINT-PIERRE, Michael KOSOROK

Dynamic Treatment Regimes



Offline Data



"The goal of the RL approach is to derive optimal DTR directly from the data" [1]

Mathematical Framework

Environment

- $t \in [0,\tau]$: **discrete time space**
- $\mathbb{S}$: State space and $s_t \in \mathbb{S}$ denotes the states of an agent at time t
- $\mathbb{A}$: Action space and $a_t \in \mathbb{A}$ denotes the chosen action of an agent at time t
- $\{\mathbb{A}(s) \mid s \in \mathbb{S}\}$: the non-empty measurable subspace of $\mathbb{A}$
- One admissible history h_t of $\mathbb{H}_t$ is $h_t = (s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t)$

Decision Process

- $(T_t)_{t\in[0,\tau]}$: the transition matrix of conditional probability transition of $\mathbb{S}$ given $\mathbb{H} \times \mathbb{A}$
- $(R_t)_{t\in[0,\tau]}$: the reward function of $\mathbb{H}_{t+1}$ in $\mathbb{R}$

Policy

A policy $\pi = (\pi_t)_{t\in\tau}$ is a sequence of conditional distribution of $\mathbb{A}$ given $\mathbb{H}_t$ such that $\forall t \in [0,\tau]$, $\forall h_t \in \mathbb{H}_t$:

$$\pi_t(\mathbb{A}(s_t) \mid h_t) = 1$$

Cumulative Reward

The long term cumulative reward at stage t is defined as $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

with $\gamma \in [0,1]$ is the discounted factor.

Optimal Policy

An optimal policy π^* is a sequence of conditional distribution such as the long term cumulated reward is maximized :

$$\pi^*(s) = \underset{\pi}{\operatorname{argmax}}\ \mathbb{E}_\nu^\pi[G_t \mid S_t = s]$$

Q-Value Based

Action-Value function : $q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$

$$\pi^*(s,a) = \underset{\pi}{\operatorname{argmax}}\ \mathbb{E}_\pi q_\pi(s,a)$$

Fitted-Q Iteration

Pseudo-Algorithm: Fitted Q-Iteration

Inputs: A set of training offline data consists of patients admissible histories h_t and their associated indexed reward r_t, $t = 0,\dots,\tau$, and a regression algorithm.

Initialization: Let $t = \tau + 1$ and $\hat{Q}_t$ be a function equal to zero everywhere on $\mathbb{S} \times \mathbb{A}$.

Iterations : Repeat computation until $t = 0$

1. $t \leftarrow t - 1$ **(Backward)**
2. Q_t is fitted with a regression algorithm though the following recursive equation :

$$Q_t(s_t, a_t) = r_t + \max_{a_{t+1}} \hat{Q}_{t+1}(s_{t+1}, a_{t+1})$$

Output: Given the sequential estimates of $\{\hat{Q}_0, \dots, \hat{Q}_\tau\}$, **the sequential optimal policies** $\{\hat{\pi}_0, \dots, \hat{\pi}_\tau\}$ can be determined.

Expert Knowledge

Assistance of a qualified professional in order to :

- Improve environmental modelling
- Incorporate observed mechanisms
- Highlight revelent decisions

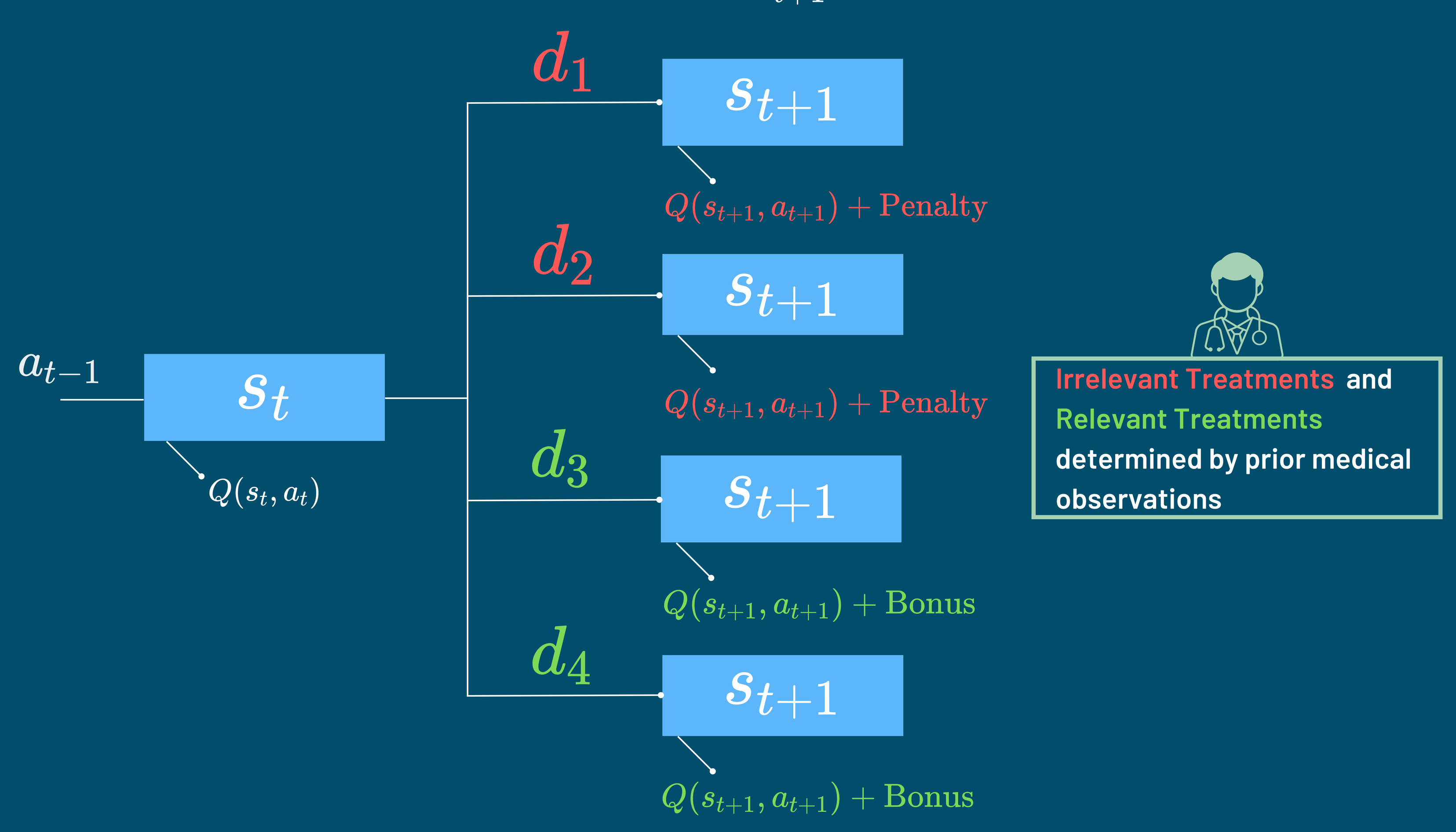
How to Integrate Expert Knowledge into Reinforcement Learning?

MODEL PREPARATION
- MODELS (SARSA, Q-LEARNING, TD…) — MIMIC-III
- LEARNING PARAMETERS — MIMIC-III
- TRANSFERT LEARNING

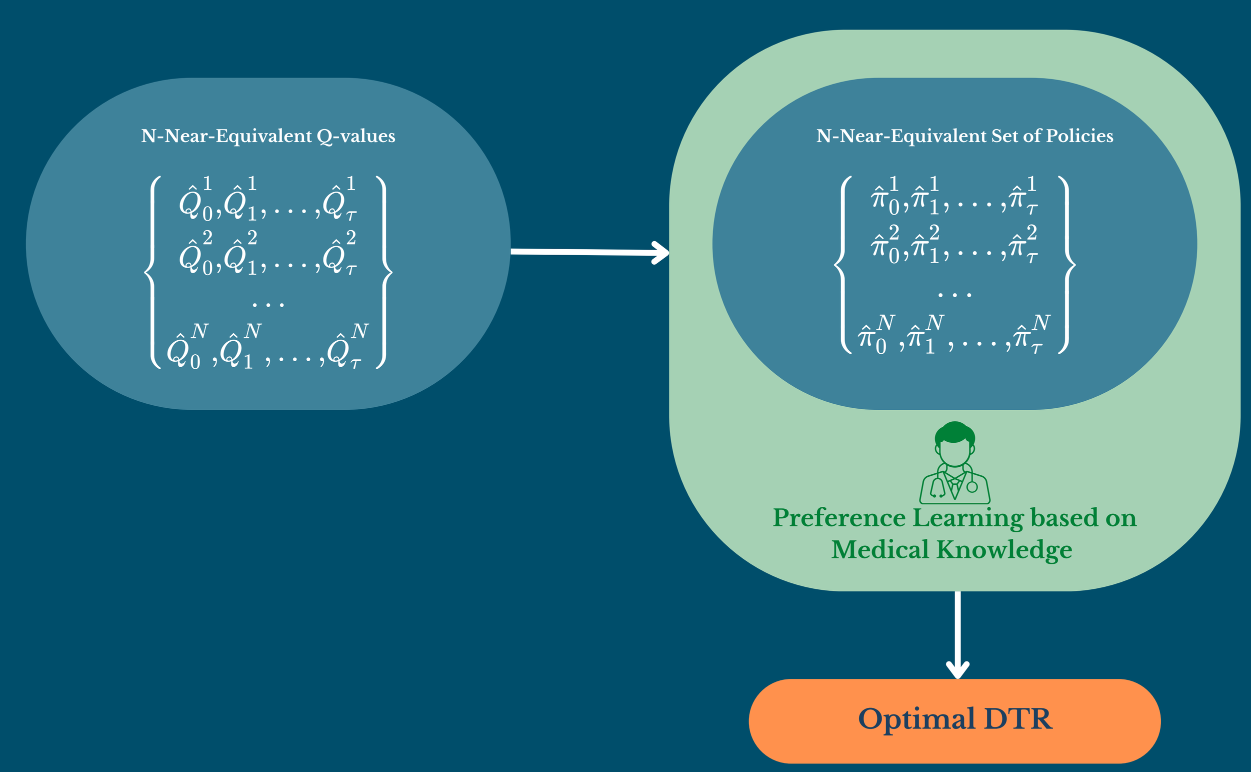
MODEL TRAINING
- INCORPORATING PRIOR MEDICAL OBSERVATION — Constraint the Q-value
- ADDING EXPLORATION — Adding H_0 hypothesis
- SUPERVISED LEARNING — Adding constraint for imitation learning

MODEL VALIDATION
- SUPERVISED LEARNING — Adding decision rule for imitation learning
- MULTIPLE POLICIES — Near-optimal set-valued policies
- HUMAIN IN THE LOOP
  - Interactive Shaping
  - Learning from categorical feedback
  - Learning from policy feedback

Constrained Q-values [2]



Near-Optimal Set-Valued Policies [3,4] and Preference Learning



- Near-equivalent actions can capture considerations such that side-effects, less invasives treatments, local availability….
- Preference Learning incorporates clinical judgments in order to rank treatments lines

References

[1] Kosorok, M. R., & Moodie, E. E. (Eds.). (2015). Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine. Society for Industrial and Applied Mathematics.

[2] Gaweda, A. E., Muezzinoglu, M. K., Aronoff, G. R., Jacobs, A. A., Zurada, J. M., & Brier, M. E. (2005, December). Incorporating prior knowledge into Q-learning for drug delivery individualization. In Fourth International Conference on Machine Learning and Applications (ICMLA'05) (pp. 6-pp). IEEE.

[3] Tang, S., Modi, A., Sjoding, M., & Wiens, J. (2020, November). Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. In International Conference on Machine Learning (pp. 9387-9396). PMLR.

[4] Lizotte, D. J., & Laber, E. B. (2016). Multi-objective Markov decision processes for data-driven decision support. The Journal of Machine Learning Research, 17(1), 7378-7405.

HEALTH YOUR IS OUR PRIORITY

Dedicated Doctor who Works Around the Clock

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

EMERGENCY SERVICES

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt.

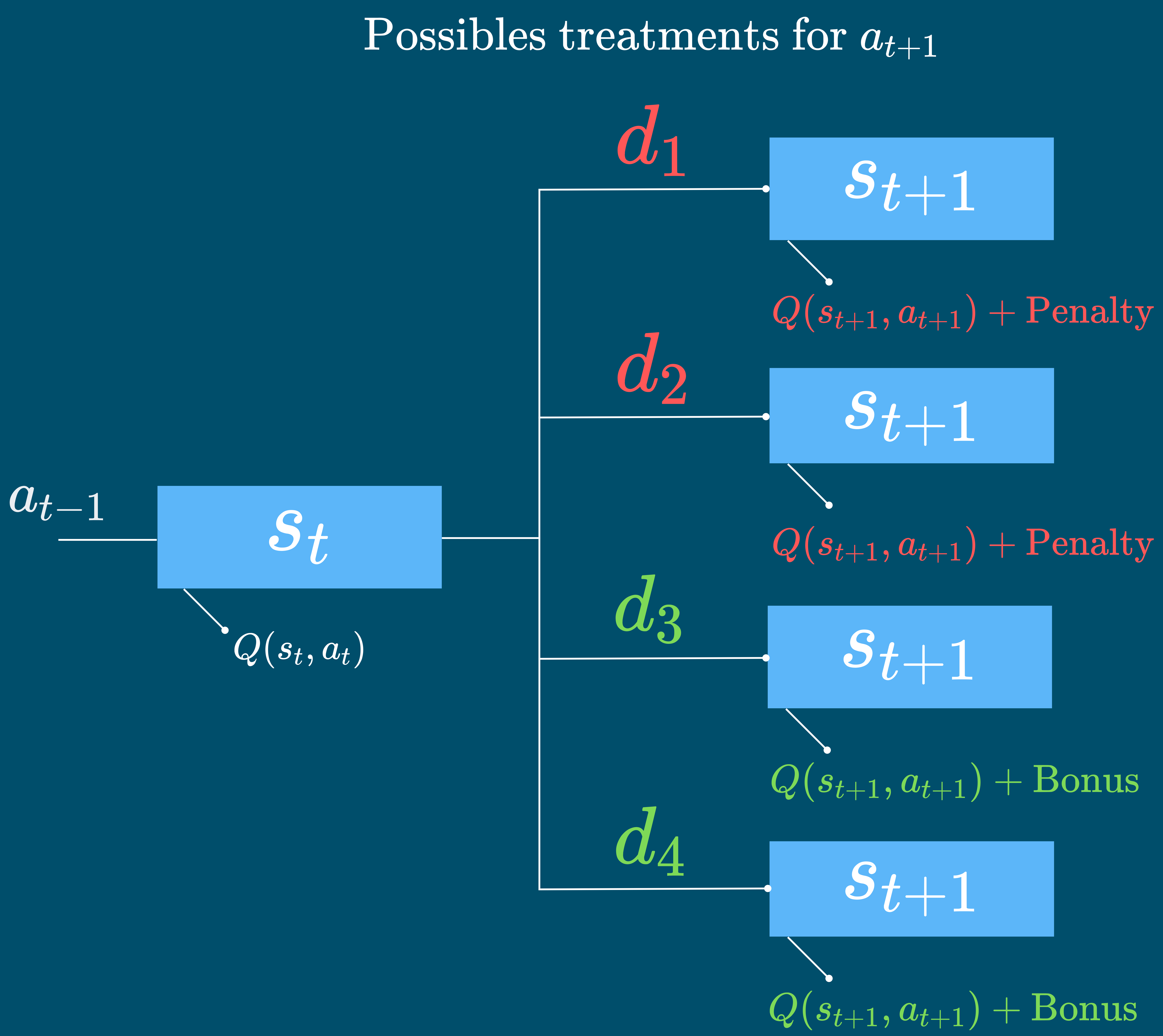
COMPLETE MEDICINE

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt.

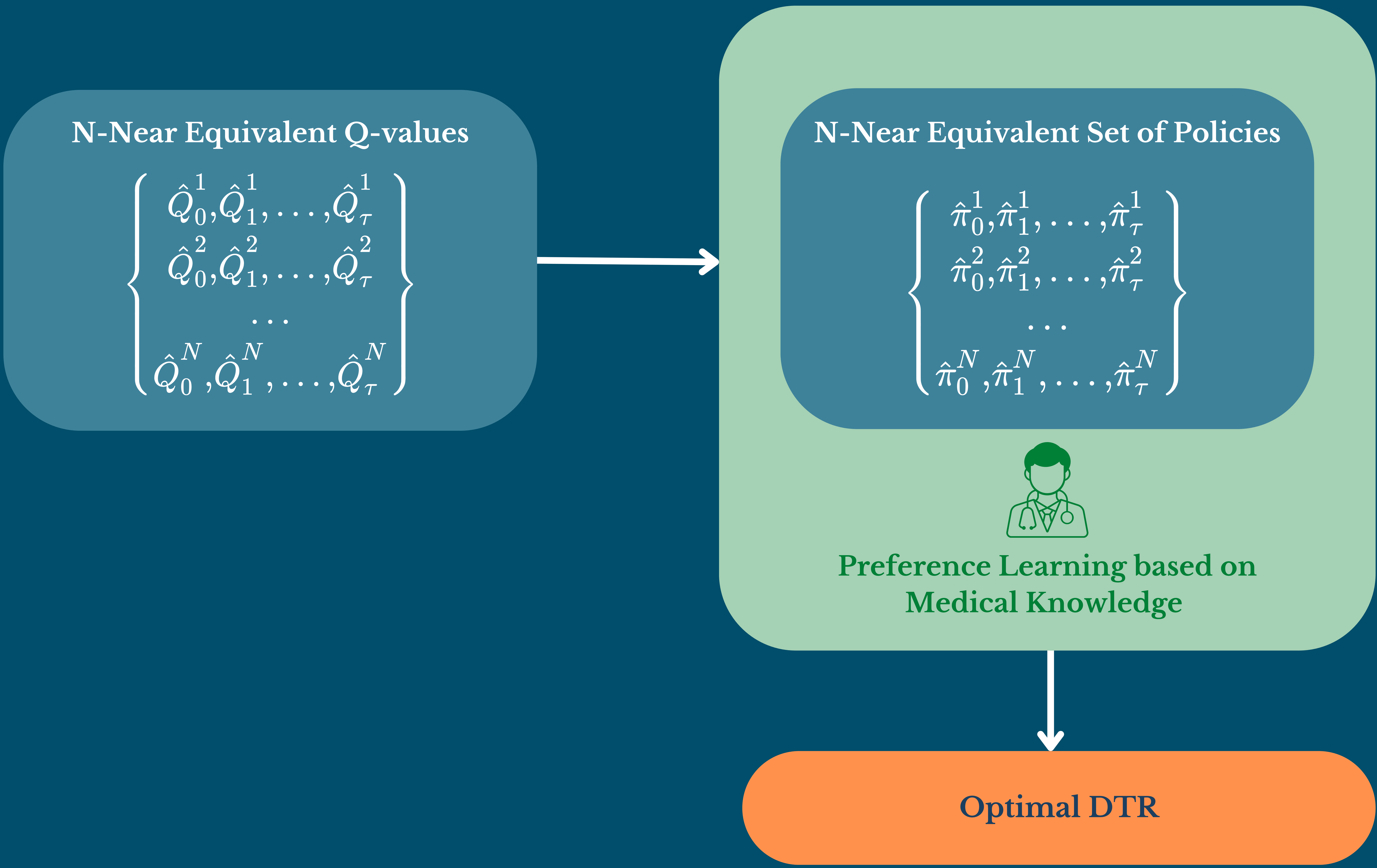
LABORATORIUM TEST

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt.

+123-456-7890 reallygreatsite.com hello@reallygreatsite.com



Irrelevant Treatments and Relevant Treatments determined by observed mechanisms


N-Near Equivalent Q-values
$\left\{ \begin{array}{c} \hat{Q}_0^1, \hat{Q}_1^1, \ldots, \hat{Q}_\tau^1 \\ \hat{Q}_0^2, \hat{Q}_1^2, \ldots, \hat{Q}_\tau^2 \\ \cdots \\ \hat{Q}_0^N, \hat{Q}_1^N, \ldots, \hat{Q}_\tau^N \end{array} \right\}$
N-Near Equivalent Set of Policies
$\left\{ \begin{array}{c} \hat{\pi}_0^1, \hat{\pi}_1^1, \ldots, \hat{\pi}_\tau^1 \\ \hat{\pi}_0^2, \hat{\pi}_1^2, \ldots, \hat{\pi}_\tau^2 \\ \cdots \\ \hat{\pi}_0^N, \hat{\pi}_1^N, \ldots, \hat{\pi}_\tau^N \end{array} \right\}$
Preference Learning based on Medical Knowledge
Optimal DTR