Artificial Intelligence-based Decision Support Systems for Precision and Digital Health

Nina Deliu[1,2,*] and Bibhas Chakraborty[3,4,5]

[1]MEMOTEF Department, Sapienza University of Rome, Italy

[2]MRC – Biostatistics Unit, University of Cambridge, UK

[3]Centre for Quantitative Medicine and the Program in Health Services and Systems Research, Duke-NUS Medical School

[4]Department of Statistics and Data Science at the National University of Singapore, Singapore

[5]Department of Biostatistics and Bioinformatics, Duke University, USA

Correspondence to: nina.deliu@uniroma1.it

Abstract

Precision health, increasingly supported by digital technologies, is a domain of research that broadens the paradigm of precision medicine, advancing everyday healthcare. This vision goes hand in hand with the groundbreaking advent of artificial intelligence (AI), which is reshaping the way we diagnose, treat, and monitor both clinical subjects and the general population. AI tools powered by machine learning have shown considerable improvements in a variety of healthcare domains. In particular, reinforcement learning (RL) holds great promise for sequential and dynamic problems such as dynamic treatment regimes and just-in-time adaptive interventions in digital health. In this work, we discuss the opportunity offered by AI, more specifically RL, to current trends in healthcare, providing a methodological survey of RL methods in the context of precision and digital health. Focusing on the area of adaptive interventions, we expand the methodological survey with illustrative case studies that used RL in real practice.

1 Introduction

In the current era of population aging and increased prevalence of chronic diseases, providing adequate health support remains one of the most urgent and complex global challenges (Prince et al., 2015). Chronic conditions such as cancer, diabetes, mental illness, or obesity tend to be of a long duration and often create a need for long-term treatment and care, carrying enormous social, medical, and economic burdens (Beaglehole et al., 2011). They are the result of a combination of genetic and physiological factors, as well as environmental and behavioral aspects that are challenging to modify, given the nature of modern lifestyle. Maintaining healthy behaviors throughout life, from diet and physical activity regimes to smoking habits, all contribute to reducing

the risk of chronic diseases, improving physical and mental capacity, and delaying care dependency.

Precision health is a relatively nascent scientific discipline that broadens the paradigm of precision medicine by including approaches that occur outside the clinical setting (Gambhir et al., 2018; Ryan et al., 2021). It seeks to develop proactive and personalized solutions to health problems, disease prevention, and health promotion. Under this framework, "disease treatment and prevention takes into account *individual variability* in genes, environment, and lifestyle for each person" (*Precision Medicine Initiative*; Collins and Varmus, 2015). It transitions from the "one-size-fits-all" standards to the formulation of treatment and prevention strategies based on the unique background and condition of each patient (Kosorok and Laber, 2019). As an illustration of this mission, a precision health system might see health proactively co-managed by healthcare providers and patients through the synchronous integration of information, starting with genotyping at birth, regular screening, and combined continuous health monitoring and provision of actionable advice and early intervention at the precise moment when the individual needs it.

Over the last decade, the precision health paradigm and healthcare in general have witnessed unprecedented innovation due to the continuous improvement and use of big data and digital technologies (Agrawal and Prabakaran, 2020). These comprise electronic tools, devices, systems, and resources that utilize increasingly fast data transmission speeds and collect, store, or process large amounts of data. Successful scientific applications of big data have already been demonstrated in numerous applications, from specific disease areas such as oncology (see e.g., *The Cancer Genome Atlas* and the *Pan-Cancer Analysis of Whole Genomes* initiatives; Tomczak et al., 2015; Aaltonen et al., 2020) or neuropsychiatry (*PsychENCODE*; PsychENCODE Consortium et al., 2015) to national initiatives. For example, the United Kingdom (UK) has now established a clear national strategy with the UK Biobank prospective cohort initiative (Allen et al., 2012), which collates together biological samples, physical measures of patient health, and sociological information such as lifestyle and demographics from 500,000 individuals. Expanding on the UK Biobank model, the American *All of US* program (All of Us Research Program Investigators et al., 2019) integrates medical records, behavioral, and family data in a unique standardized and linked database for all patients, including minorities. The goal is achieved by integrating ancillary patient data, including those collected through wearables, which are now part of daily life, interconnecting the world's population, and making health services more accessible and accountable.

The effective use of big data in healthcare is enabled by the development and deployment of artificial intelligence (AI) approaches, such as those based on machine learning (ML; Bishop, 2006). ML is a subfield of AI that uses algorithms to automatically learn from past data or experiences, making it possible to unravel patterns, associations, and causations in complex and unstructured datasets created in the era of big data (Camacho et al., 2018). In turn, it allows one to quickly provide actionable analysis on data, generating accurate prediction models–such as response of a patient to a treatment regimen–and supporting clinical practice with increasingly better decisions. An overview of successful biomedical applications using ML is provided in Deo (2015) and Rajkomar et al. (2019).

As an alternative ML area, Reinforcement Learning (RL; Sutton and Barto, 2018; Bertsekas, 2019; Sugiyama, 2015), represents a framework for interactive tasks in which the system or algorithm must learn by interacting

with the surrounding environment *sequentially*. More specifically, in RL problems, at each time step of a sequential process, an *agent* interacts with its *environment*, performs *action(s)*, and, based on a *feedback* received from the environment for the selected action(s), learns, by *trial-and-error*, on how to take better actions in order to maximize the cumulative feedback over time. This distinctive feature offers a powerful solution in a variety of healthcare domains where the problem has a sequential nature (Chakraborty and Moodie, 2013; Yu et al., 2023; Gottesman et al., 2019), such as dynamic treatment regimes (DTRs; Chakraborty and Moodie, 2013). Furthermore, the continuous improvement and use of mobile technologies has determined the development of a new area for health promotion, known as mobile health (mHealth; Istepanian et al., 2006), which aims to deliver real-time interventions tailored to individual characteristics and their rapidly changing circumstances. Such interventions are termed *just-in-time adaptive interventions* (JITAIs; Nahum-Shani et al., 2018) and have a central position in this survey. Specifically, the focus of this work is on sequential decision-making problems in healthcare and includes DTRs and JITAIs in mHealth as two key areas that have embraced the use of AI instruments.

In the coming years, AI is expected to radically transform healthcare and the way it is delivered. AI systems supported by ML have achieved considerable improvements in accuracy for diagnosis or image-based diagnosis (Myszczynska et al., 2020; McKinney et al., 2020), prognosis (Kourou et al., 2015) or drug discovery (Vamathevan et al., 2019), among others. Active research in both AI and precision health points to their convergence toward a future where healthcare is enhanced with highly personalized information and healthcare providers are empowered by decision-making support systems through augmented intelligence (Johnson et al., 2021).

Motivated by the increasing interest shown within the healthcare domain in AI technologies such as RL, this work aims to provide an overview of RL in the field of precision and digital health. Along with a survey of RL methods for specific applications in the area, we illustrate two case studies, the *PROJECT QUIT - FOREVER FREE* (Chakraborty and Moodie, 2013) and the *DIAMANTE* text messaging system (Aguilera et al., 2020; Figueroa et al., 2022), in the context of smoking cessation and physical activity, respectively, and the challenges we faced when designing the AI-based system. We believe that there is scope for important practical advances in these areas, and with this overview we aim to make it easier for methodological disciplines to join forces to assist healthcare practice and discovery and to develop the next generation of methods for AI in healthcare.

The remainder of this contribution is structured as follows. In Section 2, we provide the mathematical formalization of the general RL framework, which is further explored in Section 4.1.1 with a focus on the multi-armed bandit problem. In Section 3 and Section 4, we introduce the two areas of interest, namely DTRs and JITAIs, and extensively review existing data sources and RL methodologies for these problems. Two case studies are then illustrated in Section 3.5 and Section 4.3 for DTRs and JITAIs, respectively. Final considerations and concluding remarks are given in Section 5.

2 The Reinforcement Learning Framework

Reinforcement learning is an area of machine learning (ML) concerned with understanding how agents (e.g., systems or machines) might learn to improve their decisions through repeated experience. More formally, it aims to identify optimal decision rules (or *policies*) in *sequential decision-making problems under uncertainty* (Sutton and Barto, 2018; Bertsekas, 2019). An optimal RL policy is one that maximizes the expected long-term utility, assuming that this is likely to outweigh the associated short-term costs. The general RL framework is formalized through a continuous interaction between a *learning agent* (i.e., the decision maker) and the *environment* it wants to infer about. At each interaction stage, the agent observes some representation of the environment's *state* or *context*, and on that basis selects an *action*, that is, makes a decision. The impact of the chosen action is evaluated through a *reward* (or feedback) provided by the environment. Based on the reward received, the agent learns, by *trial-and-error*, on how to take better actions in the future to maximize the cumulative reward over time.

2.1 Basic ingredients

In reinforcement learning, differently from other ML methods, data are characterized by a sequential order and learning is carried out through many stages. For practicality, consider a discrete time space indexed by $t \in \mathbb{N} = \{0, 1, \dots, \}$. At each time t, the RL framework is described as an interaction between an agent and an unknown environment, articulated in the following three key elements:

- *State* or *context*, denoted by $X_t \in \mathcal{X}_t$, being the representation of the environment at time t. This includes the set of information (demographic and health-related covariates or physical data such as location) that may be relevant to understanding the consequences of alternative interventions.
- *Action* A_t, taken by the agent from a set of admissible actions $\mathcal{A}_t$, i.e., the set of alternative interventions. When making the choice A_t, the agent weighs the consequences of the alternatives and their likelihood, given the state X_t.
- A *reward* $Y_{t+1} \in \mathcal{Y}_{t+1} \subset \mathbb{R}$ provided by the environment in response to the chosen action A_t in correspondence with an observed state X_t. It is the information that an agent learns only after taking an action (e.g., patient response to treatment). This is closely related to the concept of utility, which should be the ultimate criterion to judge whether the entire policy works well or not.

Once action A_t is selected, together with the provision of a reward Y_{t+1}, the environment makes a transition to a new state $X_{t+1} \in \mathcal{X}_{t+1}$. Using this notation, the characterization of the RL sequential decision problem can be described as an ordered sequence or trajectory given by:

$$X_0 \to A_0 \to (Y_1, X_1) \to A_1 \to (Y_2, X_2) \to \cdots \to A_t \to (Y_{t+1}, X_{t+1}) \to \dots \tag{1}$$

In healthcare, this trajectory can be seen as the *history* of the interventions received over the course of a disease

or program, and the individual responses to treatment along with the time-varying contextual and individual health-related information.

2.2 Mathematical formalization of the general RL

Define $\mathbf{X_t} \doteq (X_\tau)_{\tau=0,\dots,t}$, $\mathbf{A_t} \doteq (A_\tau)_{\tau=0,\dots,t}$, $\mathbf{Y_{t+1}} \doteq (Y_{\tau+1})_{\tau=0,\dots,t}$, and similarly $\mathbf{x_t}$, $\mathbf{a_t}$ and $\mathbf{y_{t+1}}$, where the upper- and lower-case letters denote random variables and their particular realizations, respectively. Also define $\mathbf{H_t}$ as the *history* all the information available at time t prior to decision A_t, i.e., $\mathbf{H_t} \doteq (\mathbf{A_{t-1}}, \mathbf{X_t}, \mathbf{Y_t})$; similarly $\mathbf{h_t}$. The history $\mathbf{H_t}$ at stage t belongs to the product set $\boldsymbol{\mathcal{H}_t} = \mathcal{X}_0 \times \prod_{\tau=0}^{t-1} \mathcal{A}_\tau \times \mathcal{X}_{\tau+1} \times \mathcal{Y}_{\tau+1}$. Note that, by definition, $\mathbf{H_0} = X_0$. We assume that each longitudinal history is sampled independently according to a distribution $P_{\boldsymbol{\pi}}$, given by

$$P_{\boldsymbol{\pi}} \doteq p_0(x_0) \prod_{t \geq 0} \pi_t(a_t \mid \mathbf{h_t}) p_{t+1}(x_{t+1}, y_{t+1} \mid \mathbf{h_t}, a_t), \tag{2}$$

where:

- p_0 is the probability distribution of the initial state X_0.
- $\boldsymbol{\pi} \doteq \{\pi_t\}_{t \geq 0}$ represents the agent's *policy* and determines the sequence of actions generated throughout the decision-making process. More specifically, π_t maps histories of length t, $\mathbf{h_t}$, to a probability distribution over the action space $\mathcal{A}_t$, i.e., $\pi_t(\cdot \mid \mathbf{h_t})$. The conditioning symbol "|" in $\pi_t(\cdot \mid \mathbf{h_t})$ reminds us that the policy defines a probability distribution over $\mathcal{A}_t$ for each $\mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}$. Sometimes, A_t is uniquely determined by the history $\mathbf{H_t}$, therefore the policy is simply a function of the form $\pi_t(\mathbf{h_t}) = a_t$. We call it *deterministic policy*, in contrast with *stochastic policies* that determine actions probabilistically.
- $\{p_t\}_{t \geq 1}$ are the unknown *transition probability distributions* that characterize the dynamics of the environment. At each time $t \in \mathbb{N}$, the transition probability p_t assigns to each trajectory $(\mathbf{x_{t-1}}, \mathbf{a_{t-1}}, \mathbf{y_{t-1}}) = (\mathbf{h_{t-1}}, a_{t-1})$ at time $t-1$ a probability measure over $\mathcal{X}_t \times \mathcal{Y}_t$, specifying the probability of transition to new states in $\mathcal{X}_t$ with reward in $\mathcal{Y}_t$, from history $\mathbf{h_{t-1}}$ and action a_{t-1}, i.e., $p_t(\cdot, \cdot \mid \mathbf{h_{t-1}}, a_{t-1})$.

At each time t, the transition probability distribution $p_{t+1}(x_{t+1}, y_{t+1} \mid \mathbf{h_t}, a_t)$ gives rise to: (i) the *state-transition probability distribution* $p_{t+1}(x_{t+1} \mid \mathbf{h_t}, a_t)$, i.e., the probability of transitioning to state x_{t+1} having observed a history $\mathbf{h_t}$ and taking action a_t; and (ii) the *immediate reward distribution* $r_{t+1}(y_{t+1} \mid \mathbf{h_t}, a_t, x_{t+1})$, which specifies the reward Y_{t+1} after transitioning from a history $\mathbf{h_t}$ to x_{t+1} under action a_t. To better incorporate uncertainty, we assume a stochastic reward distribution. An illustrative representation of the RL framework is provided in Figure 1. The cumulative discounted sum of immediate rewards from time t onward is known as *return*, say $\mathbf{R_t}$, and is given by

$$\mathbf{R_t} \doteq Y_{t+1} + \gamma Y_{t+2} + \gamma^2 Y_{t+3} + \cdots = \sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1}, \quad t \in \mathbb{N}. \tag{3}$$

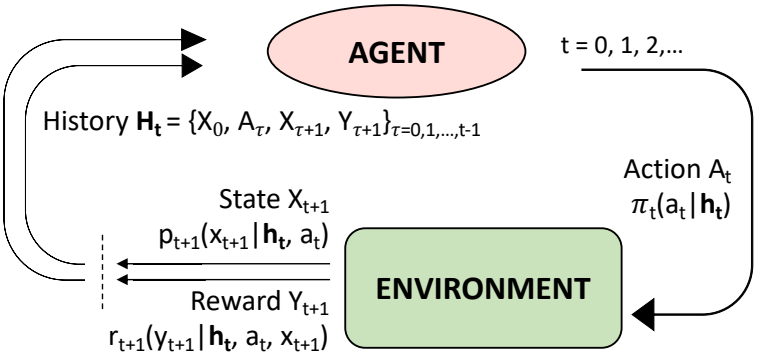

Figure 1: Illustration of the general RL framework.

The *discount rate* $\gamma \in [0, 1]$ determines the current value of future rewards: a reward received τ time steps in the future is worth only γ^τ times what it would be worth if it were received immediately. If $\gamma = 0$, the agent is *myopic* in being concerned only with maximizing the immediate reward, that is, $\mathbf{R_t} = Y_{t+1}$, with the convention that $0^0 = 1$. If $\gamma = 1$, the return is *undiscounted* and it is well defined (finite) as long as the time-horizon is finite, i.e., $T < \infty$ (Sutton and Barto, 2018).

Remark: On-policy Vs. off-policy learning The agent's policy $\boldsymbol{\pi}$ determines the sequential selection of actions. In a randomized experiment context, for example, it may define the randomization probabilities of each intervention at each decision point t. In such a setting, an agent can be interested in learning and optimizing the policy $\boldsymbol{\pi}$ while following it, that is, from experiences sampled directly from $\boldsymbol{\pi}$. This type of learning is termed *on-policy* or *online* learning, and policy $\boldsymbol{\pi}$ represents both the *exploration policy* (the one that generates the data) and the *target policy* (the one we learn about). In contrast, there are settings where the agent learns from previously collected data without interacting with the environment to collect samples (e.g., observational data). In this type of learning, termed *off-policy* or *offline*, we say that the target policy is learned from data "off" the target policy, which are determined according to a policy $\boldsymbol{\pi}$ that can be either the exploration policy (when known, e.g., in randomized studies), or, more generally, an observed or *behavior policy*. Similar concepts are used in statistical and causal inference for referring to the estimation of unknown quantities of interest such as parameters.

Taking into account a potential misalignment between the target policy, say $\mathbf{d}$, and the one used to generate the data $\boldsymbol{\pi}$ (either exploration or behavior), the RL problem at any time t is to learn an optimal way to choose the set of actions, i.e., an *optimal policy* $\mathbf{d_t^*} \doteq \{d_t^*\}_{\tau \geq t}$, so as to maximize the expected future return. Formally,

$$\boldsymbol{d_t^*} = \arg\max_{\boldsymbol{d_t}} \mathbb{E}_{\boldsymbol{d}}[\mathbf{R_t}] = \arg\max_{\boldsymbol{d_t}} \mathbb{E}_{\boldsymbol{d}} \left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1} \right], \tag{4}$$

where the expectation is meant with respect to a trajectory distribution $P_{\mathbf{d}}$ analogous to Eq. (2), where the policy $\boldsymbol{\pi}$ that generated the data is replaced by the target policy $\mathbf{d}$ we want to learn about. Note that the *expected* return is the most common approach to handle decision making under uncertainty (De Lara et al., 2008).

For learning optimal policies, various methods have been developed so far in the RL literature: see Sutton and Barto (2018) and Sugiyama (2015) for an overview. A traditional approach is through *value functions*, which define a partial ordering over policies, with insightful information on the optimal ones. In fact, optimal policies share the same (optimal) value function, and comparing estimated value functions of different candidate policies offers a way to understand which strategy may offer the greatest expected outcome.

There are two types of value functions: i) *state-value* or simply *value* functions, say $V_t^{\boldsymbol{d}}$, representing how good it is for an agent to be in a given state, and ii) *action-value* functions, say $Q_t^{\boldsymbol{d}}$, indicating how good it is for the agent to perform a given action in a given state. These are formally defined as:

$$V_t^{\boldsymbol{d}}(\mathbf{h_t}) \doteq \mathbb{E}_{\boldsymbol{d}}\left[\mathbf{R_t}|\mathbf{H_t}=\mathbf{h_t}\right] = \mathbb{E}_{\boldsymbol{d}}\left[\sum_{\tau\geq t}\gamma^{\tau-t}Y_{\tau+1}\middle|\mathbf{H_t}=\mathbf{h_t}\right], \tag{5}$$

$$Q_t^{\boldsymbol{d}}(\mathbf{h_t}, a_t) \doteq \mathbb{E}_{\boldsymbol{d}}\left[\mathbf{R_t}|\mathbf{H_t}=\mathbf{h_t}, A_t=a_t\right] = \mathbb{E}_{\boldsymbol{d}}\left[\sum_{\tau\geq t}\gamma^{\tau-t}Y_{\tau+1}\middle|\mathbf{H_t}=\mathbf{h_t}, A_t=a_t\right], \tag{6}$$

$\forall t \in \mathbb{N}$, $\forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}$ and $\forall a_t \in \mathcal{A}_t$, with $\mathbf{H_t}$ and A_t such that $\mathbb{P}(\mathbf{H_t}=\mathbf{h_t})>0$ and $\mathbb{P}(A_t=a_t)>0$. By definition, at stage $t=0$, $V_0^{\boldsymbol{\pi}}(\mathbf{h_0}) \doteq V_0^{\boldsymbol{\pi}}(x_0)$; while for the terminal stage, if any, the state-value function is 0.

At stage t, the *optimal value function* $V_t^{\boldsymbol{d}^*}$ yields the largest expected return for each history, and the *optimal Q-function* $Q_t^{\boldsymbol{d}^*}$ yields the largest expected return for each history-action pair, i.e.,

$$Q_t^*(\mathbf{h_t}, a_t) \doteq \max_{\boldsymbol{d_t}} Q_t^{\boldsymbol{d}}(\mathbf{h_t}, a_t), \quad \forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}, \forall a_t \in \mathcal{A}_t. \tag{7}$$

A fundamental property of the value functions used throughout RL is that they satisfy particular recursive relationships, known as *Bellman* equations (Bellman, 1957). In the Q-value case, for instance, for any policy $\mathbf{d}$, the following consistency condition, expressing the relationship between the quality of an history-action and the quality of the successors, holds:

$$Q_t^*(\mathbf{h_t}, a_t) = \mathbb{E}\Big[Y_{t+1} + \gamma \max_{a_{t+1}\in\mathcal{A}_{t+1}} Q_{t+1}^*(\mathbf{h_{t+1}}, a_{t+1}) \mid \mathbf{H_t}=\mathbf{h_t}, A_t=a_t\Big], \tag{8}$$

$\forall a_t \in \mathcal{A}_t, \forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}, \forall t \in \mathbb{N}$, and with discrete state and action spaces. Here, the expectation $\mathbb{E}$ is taken with respect to the transition distribution p_{t+1} only, which does not depend on the policy; thus, the subscript $\boldsymbol{d}$ can be omitted.

The property in Eq. (8) allows estimation of (optimal) value functions recursively, from T backward in time. In finite-horizon *dynamic programming* (DP), this technique is known as *backward induction* and represents one of the main methods to solve the Bellman equation. In infinite- and indefinite-horizon problems, traditional backward induction is not possible, given the impossibility of extrapolating beyond the time horizon in the observed data. To overcome this issue, alternative methods and additional assumptions (e.g., discounting and boundedness of rewards) are typically taken into account. Common strategies (e.g., V-learning, which we review in Section 3.3; Luckett et al., 2020), focuses on time-homogeneous Markov processes.

Due to its generality, RL is studied and employed in many disciplines, from game theory to education and healthcare. In this work, our goal is to review common RL classes that have been studied or used to support decision making in precision and digital health. In particular, we cover the RL methods studied in the DTR literature (e.g., Q-learning and outcome weighted learning) and in digital health (with a main interest in the multi-armed bandit class, which we discuss in Section 4.1.1). For readers interested in the general area of RL and the broad spectrum of existing methods, without a specific application in mind, we refer to Sutton and Barto (2018); Sugiyama (2015).

3 Dynamic Treatment Regimes in Precision Health

Clinical or behavioral treatments often involve a series of decisions over time that account for the continuously evolving histories of individuals. For example, weight loss management involves a sequence of decisions at multiple stages of weight progression. Initially, individuals affected by excess body weight undergo lifestyle modifications (such as diet and exercise), and based on their body mass index (BMI), may be treated with pharmacologic therapies to achieve the desired body weight. Then, if the individual *responds* (i.e., shows a significant weight loss), the physician may prescribe a maintenance therapy (typically diet and exercise) to maintain weight at a reduced level. Otherwise, the clinician prescribes a *second-line* therapy, to try to induce body weight reduction. There exist many possible therapies. The aim of the physician is to choose the sequence of therapies that leads to the best possible outcome, e.g., long-term maintenance of lost weight, for that individual.

Similarly, the treatment of cancer, diabetes, mental health disorders, or the management of addiction problems requires a series of decisions by which the physician can start, stop, maintain, modify, or adjust interventions based on the patient's response and evolving characteristics. This sequence of decisions constitutes a *dynamic treatment regime* or *regimen* (Murphy, 2003; Chakraborty and Moodie, 2013), alternatively known as adaptive interventions or strategies (Collins et al., 2004; Lavori and Dawson, 2000).

Dynamic treatment regimes (DTRs) offer a vehicle to operationalize the sequential decision-making process involved in clinical practice and can also be viewed as a *decision support system*. A DTR is defined as a sequence of decision rules, one per stage of intervention, dictating how to personalize treatments to patients based on their baseline and evolving history (*time-varying, dynamic state*), repeatedly adjusting over time in response to ongoing performance (Almirall et al., 2014; Nahum-Shani et al., 2018). Thus, the treatment regime is "dynamic" within a person over time, varying because the person or disease is changing, with the goal of obtaining the best results for that individual.

The existing DTR frameworks (Collins et al., 2004; Almirall et al., 2014) highlight four components that play an important role in the design of these interventions:

(i) The critical **decision points**, specifying the time points at which a decision concerning intervention (e.g., continue, alter, add, or subtract treatment) has to be made; here we assume a finite or countable number of times $t = 0, 1, \ldots$;

(ii) The **decisions or treatment options** at each time t, denoted by $A_t \in \mathcal{A}_t$, where $\mathcal{A}_t$ is the decision or action space, generally discrete;

(iii) The **tailoring variable(s)** at each time t, say $X_t \in \mathcal{X}_t$, with $\mathcal{X}_t \subseteq \mathbb{R}^p$, capturing individuals' baseline and time-varying information for personalizing decision-making;

(iv) The **decision rules** $\mathbf{d} = \{d_t\}_{t\geq 0}$, that, at each time t, link the tailoring variable(s) to specific decisions.

Treatment options $A_t \in \mathcal{A}_t$ are not limited to different medications or drugs, but can also include different dosages (duration, frequency or amount; Voils et al., 2012; Chen et al., 2016), various tactical options (for example, increase, change, maintain), modes of administration (for example, oral or injection), timing schedules (Nie et al., 2021), behavioral interventions, or no further treatment. Tailoring variables $X_t \in \mathcal{X}_t$ refer to patient and treatment information available up to the time of the critical decision, and may include previous treatment and disease history, genetic information, diagnostic test results, etc. Once the four elements are defined, each decision rule $\mathbf{d} = \{d_t\}_{t\geq 0}$ takes the individual characteristics $X_t \in \mathcal{X}_t$ of a subject and their treatment history observed up to that stage $\{A_t\}_{t=0,1,\ldots,t-1}$ as input and outputs a recommended treatment strategy at that stage. The dynamic treatment regime $\mathbf{d_t} = (d_0, \ldots, d_t)$ is regarded a multistage regime with each d_τ, $\tau = 0, \ldots, t$ being a mapping of the entire evolving history $\mathcal{X}_0 \times \mathcal{A}_0 \times \cdots \times \mathcal{A}_{\tau-1} \times \mathcal{X}_\tau$ to $\mathcal{A}_\tau$. Unlike average-based single-stage protocols, DTRs explicitly incorporate the heterogeneity in treatment effect among individuals and *across time* within an individual. As such, it provides an attractive framework for personalized treatments in longitudinal settings. Furthermore, by treating only those who show a need for treatment, DTRs hold the promise of reducing noncompliance due to overtreatment or undertreatment (Lavori and Dawson, 2000).

3.1 RL methods for constructing optimal DTRs

One of the main research goals in the field of personalized dynamic treatments is to construct *optimal* DTRs, that is, to identify the treatment rule(s) that result in the best (typically long-term) mean outcome, i.e., with the highest utility. Most attempts to achieve this goal essentially require knowing or estimating the prespecified *utility* function or some variations of it. For example, Murphy (2003) defines *regret* (i.e., *loss*) functions, while Robins (2004) introduces *blip* functions (Kitagawa and Tetenov, 2018), alternatively known as *welfare gains* in econometrics.

Methodologies for estimating optimal DTRs are of considerable interest within the domain of precision health and comprise a growing body of research in both computer science and statistics (Chakraborty and Moodie, 2013). On the one hand, the sequential decision-making nature of DTR problems perfectly conforms to the RL framework, thus attracting increasing attention in the ML literature. On the other hand, the need to quantify causal relationships, rather than mere associations, called for the intervention of the causal inference community. Since the underlying system dynamics is often unknown, inferring the consequences of executing a policy $\mathbf{d} = \{d_t\}_{t\geq 1}$ and understanding the causal effects on an outcome is a challenging task. We refer to Deliu and Chakraborty (2022) and Tsiatis et al. (2021) for the broad range of aspects related to DTR (including

inference), while here we focus exclusively on the role of AI, more specifically RL methods, in deriving optimal DTRs. Notably, due to the similarity between the two problems and their components, RL represents one of the main approaches employed in the DTR literature. A preliminary non-exhaustive correspondence table between the RL and DTR terminologies is reported in **Table 1**.

Table 1: Terminology correspondence between reinforcement learning (RL) and dynamic treatment regimes (DTRs).

Notation	**Terminology**	
	RL	DTRs
i	Trajectory, Unit	Patient, Subject, Individual
t	Time, Step, Round	Stage, Interval, Round
X	State, Context	Covariates
A	Action, Arm	Treatment, Intervention
Y	Reward, Feedback	Outcome, Response
$\mathbf{H}$	History, Filtration	Time Varying History
$\boldsymbol{\pi}/\mathbf{d}$	Policy	Dynamic Treatment Regime(n), Adaptive Intervention

Most of the existing work in DTRs rely on the finite-horizon setting (with a prespecified horizon $T < \infty$), and the strongly connected *offline learning* procedures, wherein estimation is based on existing longitudinal data. Recent solutions in the indefinite-horizon setting, particularly suitable for electronic health record (EHR) data and for chronic diseases–where the number of stages can be arbitrarily large and are not known a-priori–are proposed in Ertefaie and Strawderman (2018) and Luckett et al. (2020).

The fundamental learning mechanism for deriving optimal policies consists in two approaches: *direct* and *indirect* methods. Direct methods learn optimal policies by directly looking for the policy that maximizes an objective (typically the expected return or value function) within a class of policies. On the contrary, indirect methods attempt first to estimate a value function and then to determine an optimal policy based on the learned value function. In the RL literature, direct and indirect methods are sometimes referred to as *model-free* and *model-based* algorithms (Sutton and Barto, 2018). However, more subtle classifications (see e.g., Guan et al., 2021; Sugiyama, 2015) tend to make a clearer separation between the two categories in the sense that direct/indirect are used for the learning process, while model-free/model-based refer to the modeling assumptions for the environment. A graphical illustration is provided in Figure 2. In what follows, we review existing RL

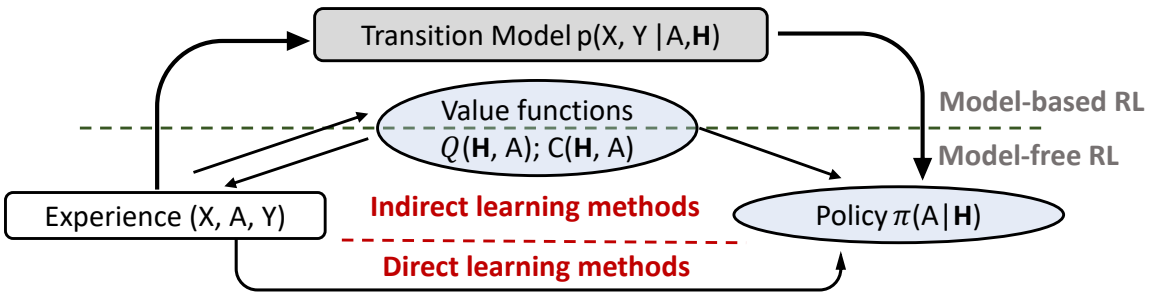

Figure 2: Illustrative comparison of direct vs indirect RL methods.

techniques for developing DTRs, covering both finite- and indefinite-horizon settings, and adopting the direct vs. indirect taxonomy, in line with the current DTR literature (Chakraborty and Murphy, 2014; Deliu and Chakraborty, 2022).

Remark: Causal inference For simplicity of exposition, the illustration of this subject will be carried out in a simplified framework where the main assumptions of causal inference (see e.g., Chakraborty and Murphy, 2014) hold. It follows that RL can operate in a simplified causal inference problem in which actions are unconfounded. For a comprehensive treatment of causal inference, we refer to Hernan and Robins (2023), while for a specific characterization of the framework for the DTR problem, we refer to Chakraborty and Murphy (2014); Deliu and Chakraborty (2022).

3.2 Indirect RL methods in DTRs

Indirect methods focus on estimating an optimal objective function (typically, an expectation of the outcome variable such as the Q-function presented in Eq. (6)), and then obtaining the associated policy. These methods are mainly based on iterative techniques such as dynamic programming (DP) and approximate dynamic programming (ADP), and include the Q-learning (Murphy, 2005b) approach that we illustrate below.

We mainly focus on the finite-horizon setting, where the utility function is optimized over a fixed and prespecified period of time T, and, for the sake of simplicity, we consider deterministic policies which map histories $\mathbf{h}$ directly into actions or decisions, that is, $\mathbf{d}(\mathbf{h}) = \mathbf{a}$.

Q-learning Q-learning (Watkins, 1989) represents the core of modern RL and one of the most popular strategies in DTR research. Its fundamental idea is based on iterative improvement of the estimates of the Q-function at a given stage t, starting from a previous estimate and following the Bellman rule in Eq. (8). That is,

$$Q_t(\mathbf{h_t}, a_t) \leftarrow Q_t(\mathbf{h_t}, a_t) + \alpha_t \Big[Y_{t+1} + \gamma \max_{a_{t+1} \in \mathcal{A}_{t+1}} Q_{t+1}(\mathbf{h_{t+1}}, a_{t+1}) - Q_t(\mathbf{h_t}, a_t) \Big].$$

The constant α_t determines to what extent the newly acquired information should override the old information, i.e., how fast learning takes place: a factor of 0 will make the learner not learn anything, while a factor of 1 would make the learner fully update based on the most recent information. The discount factor γ balances the immediate rewards of the learner with future rewards, and in a finite-horizon problem it is generally set to one.

The original version of this approach is known as tabular Q-learning, and it is based on storing the Q-function values for each possible state and action in a lookup table and choosing the one with the highest value. Under some appropriate and rigorous assumptions (Watkins, 1989), Q_t has been shown to converge to the optimal Q-function Q_t^* with probability 1. However, this procedure is practical for a small number of problems because it can require many thousands of training iterations to converge. In addition, it represents value functions in arrays or tables, based on each state and action. Thus, large state spaces lead not just to memory issues for large tables but also to time problems needed to fill them accurately. A more recent version of Q-learning, known as *Q-learning with function approximation* (FA), offers a powerful and scalable tool to overcome both the modeling requirements and the computational burden to solve an RL problem through backward induction.

The main idea of Q-learning with FA is first to estimate the Q-functions using an approximator, e.g., regression models, neural networks or decision trees, and then to derive the estimated policy based on the

estimated Q-functions. Considering an approximation space for each of the T stage-specific Q-functions, e.g., $\mathcal{Q}_t \doteq \{Q_t(\mathbf{h_t}, a_t; \theta_t) : \theta_t \in \Theta_t\}$, with Θ_t the parameter space, an optimal stage-t policy estimate is given by:

$$\hat{d}_t^*(\mathbf{h_t}) = \arg\max_{a_t \in \mathcal{A}_t} \hat{Q}_t^*(\mathbf{h_t}, a_t) \doteq \arg\max_{a_t \in \mathcal{A}_t} Q_t^*(\mathbf{h_t}, a_t; \hat{\theta}_t) \doteq d_t^*(\mathbf{h_t}; \hat{\theta}_t)), \quad t = 0, \ldots, T.$$

An optimal regime $\hat{\mathbf{d}}^* = (d_1^*(x_1; \hat{\theta}_1), d_2^*(\mathbf{h_2}; \hat{\theta}_2), \ldots, d_T^*(\mathbf{h_T}; \hat{\theta}_T))$ is obtained by following Bellman's optimality equation in Eq. (8), and by recursively estimating Q_t^* backward in time $t = T, T-1, \ldots, 1$. Noticing that Q-functions are conditional expectations, regression models represent a natural approach. For the complete general iterative procedure, as well as more specific examples, we point to Deliu and Chakraborty (2022). By using generalized linear models, one may extend the Q-learning method to binary and count outcomes, and an accelerated failure time model can be incorporated for survival outcomes. In the DTR arena, Q-learning generalizations to diverse outcomes have been implemented for censored data Goldberg and Kosorok (2012); Zhao et al. (2011, 2020), binary data Moodie et al. (2014), or composite measures attempting to balance different objectives (Laber et al., 2014), among others.

In order for $\hat{\mathbf{d}}^*$ to be a consistent estimator of the true optimal regime $\mathbf{d}^*$, it is important to recognize that all the models for the Q-functions should be correctly specified (Schulte et al., 2014). To address this problem, several FA alternatives such as *support vector regression* and *extremely randomized trees* (Zhao et al., 2009), or *deep neural networks* (Liu et al., 2017; Atan et al., 2018; Raghu et al., 2017) have been proposed. We now illustrate the latter, given the attention it has attracted in recent years.

Deep Q-learning. The success achieved by Q-learning in many complex domains has been largely enabled by the use of advanced FA techniques such as *deep neural networks* (Mnih et al., 2015). We call this approach *Deep Q-learning* (DQL). In DQL, a neural network (Goodfellow et al., 2016) is used to approximate the Q-function. More specifically, at each time t, a DNN is used to fit a model for the Q-function in a supervised way. States and actions $\{(\mathbf{H}_{t,i}, A_{t,i})\}_{i=1,\ldots,N}$ are given as inputs (in the input layer), and the Q-values of all possible actions are generated as outputs $\{Q_t(\mathbf{H}_{t,i}, A_{t,i}; \hat{\mathbf{W}}, \hat{\mathbf{b}})\}_{i=1,\ldots,N}$ (in the output layer), leading to a labeled set of data $\{(\mathbf{H}_{t,i}, A_{t,i}), Q_t(\mathbf{H}_{t,i}, A_{t,i}; \hat{\mathbf{W}}, \hat{\mathbf{b}})\}_{i=1,\ldots,N}$. Input data are non-linearly transformed based on the unknown weight W and bias b parameters and carried out through the neurons of the hidden layers. Figure 3 shows a schematic of a *feed-forward neural network* used within RL. It is characterized by a set of neurons, structured in layers, where each neuron processes the information from one layer to the next. The collected data are stored and used for continuously updating the Q-function parameter estimates. To allow exploration, each decision is determined by an exploration scheme (typically ϵ-greedy) that probabilistically chooses between the action with the highest Q-value and a random action.

Within the DTR literature, DQL has been implemented with EHR data in Liu et al. (2017) and Raghu et al. (2017) for graft-versus-host disease and sepsis treatment, respectively. Compared to its shallow counterpart, the DQL framework is particularly suitable for (i) automatically extracting and organizing discriminative information from the data and (ii) exploring the high-dimensional action and state spaces and making personalized treatment

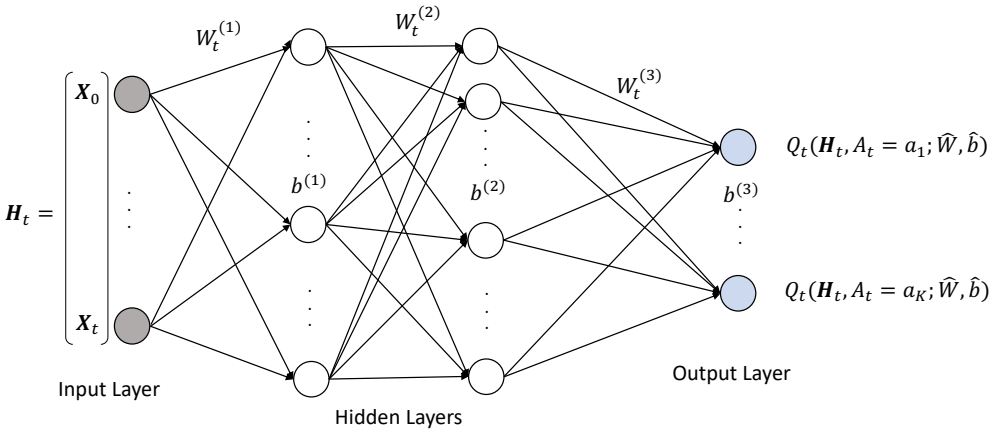

Figure 3: Representation of a feed-forward neural network with four layers used within Q-learning.

recommendations.

3.3 Direct RL methods in DTRs

Direct methods, also known in the RL literature as direct policy search methods (Ng and Russell, 2000), seek to maximize the return by learning the optimal policy directly, without involving the estimation of intermediate quantities such as optimal Q-functions or contrasts. These methods typically do not assume models for conditional mean outcomes; thus, they are referred to as "nonparametric". However, they may consider a parameterization for the policies or regimes class.

In direct methods, a class of policies $\mathcal{D}$, often indexed by a parameter, say $\psi \in \Psi$, is first prespecified. Then, for each candidate regime $d \in \mathcal{D}$, an estimate of the corresponding *utility* is obtained. The utility can be a summary of one outcome, such as percent days of abstinence in an alcohol dependence study or a composite outcome. For example, in Wang et al. (2012) the utility is a compound score that combines information on treatment efficacy, toxicity, and risk of disease progression. Here, without loss of generality, we take the utility to be the value of the policy; see Eq. (5). The regime in $\mathcal{D}$ that maximizes the value function is the estimated optimal DTR, that is, $\hat{d}^* \doteq \arg\max_{d\in\mathcal{D}} \hat{V}_d$, or $\hat{d}^* \doteq \arg\max_{\psi\in\Psi} \hat{V}_{d_\psi}$ for parametric classes. A common example of parametric classes is the *soft-max* class $\mathcal{D} \doteq \{\pi(a_k|\mathbf{x},\psi) = e^{-\mathbf{x}^T\psi_k}/\sum_{j=1}^K e^{-\mathbf{x}^T\psi_j} : \psi \in \Psi, k = 1,\ldots,K\}$, where $a_1,\ldots,a_K$ are the K possible treatments and $\psi \doteq (\psi_1^T,\ldots,\psi_K^T)$ is the vector of parameters for the K treatments indexing the class of policies.

Most statistical work in this area is based on the *inverse probability of treatment weighting* (IPTW) estimator (Robins, 1994), used for estimating value functions (Zhang et al., 2012, 2013), in classification-based frameworks such as *outcome weighted learning* (OWL; Zhao et al., 2012, 2015; Liu et al., 2018), and in combination with ML approaches such as decision trees (Laber and Zhao, 2015; Tao et al., 2018). Particularly useful in observational data, where the exploration and target policies differ, IPTW (Robins, 2000) makes use of importance sampling to change the distribution under which the regime's value is computed. In doing so, assuming that P_d is absolutely continuous with respect to P_π, it basically weights outcomes according to the

relative probability of interventions occurring under the target $\mathbf{d}$ and exploration $\boldsymbol{\pi}$ policies. The value function then can be rewritten as:

$$V^d = \mathbb{E}_d[Y] = \int Y dP_d = \int Y \left(\frac{dP_d}{dP_\pi}\right) dP_\pi = \tag{9}$$
$$= \int \left(\prod_{t=0}^{T} \frac{\mathbb{I}[A_t = d_t(H_t)]}{\pi_t(A_t|H_t)}\right) Y dP_\pi \doteq \int w_{d,\pi} Y dP_\pi.$$

To estimate V^d, the Monte Carlo (MC) estimator given by $\hat{V}^d \doteq \mathbb{P}_N [w_{d,\pi} Y]$, where $\mathbb{P}_N$ denotes the empirical average over N trajectories, is generally used. By the Strong Low of Large Numbers, the MC estimator is unbiased, but its variance is unbounded. To stabilize this estimator, the weights $w_{d,\pi}$ are normalized by their sample mean, leading to the IPTW estimator:

$$\hat{V}^d_{IPTW} \doteq \frac{\mathbb{P}_N [w_{d,\pi} Y]}{\mathbb{P}_N [w_{d,\pi}]}. \tag{10}$$

The technique also allows balancing the confounders across levels of treatment: the higher the probability of receiving a specific treatment conditioned on the confounder X, $\pi(A|X)$, the lower the weight $w_\pi = 1/\pi(A|X)$ of their outcome Y.

When π is known (e.g., randomized trials), the IPTW estimator is consistent, but it can be highly variable due to the presence of nonsmooth indicator functions inside the weights. An alternative version, which integrates the properties of the IPTW estimator with those of regression–assuming models for both the propensity score and the (conditional) mean outcome–is the *augmented inverse probability of treatment weighting* (AIPTW) estimator (Zhang et al., 2012). Assume a single-stage treatment regime with two treatment options ($A \in \{a, a'\}$), and let $H = X_0$ to be a patient's history, $d(H) \doteq d(H;\psi)$ a treatment regime indexed by ψ, $\mu(A, H; \hat{\beta})$ an estimated model for the mean outcome $\mathbb{E}[Y|H, A]$, and $\pi(A|H, \hat{\gamma})$ an estimated propensity score. Then, the AIPTW estimator is defined as:

$$\hat{V}^d_{AIPW} \doteq \mathbb{P}_N \left\{ \frac{\mathbb{I}[A = d(H;\psi)]Y}{\pi(H;\psi,\hat{\gamma})} - \frac{\mathbb{I}[A = d(H;\psi)] - \pi(H;\psi,\hat{\gamma})}{\pi(H;\psi,\hat{\gamma})} \times \mu(H;\psi,\hat{\beta}) \right\},$$

where,

$$\pi(H;\psi,\hat{\gamma}) \doteq \pi(a|H,\hat{\gamma})\mathbb{I}[d(H;\psi) = a] + \pi(a'|H,\hat{\gamma})\mathbb{I}[d(H;\psi) = a'],$$
$$\mu(H;\psi,\hat{\beta}) \doteq \mu(a, H;\hat{\beta})\mathbb{I}[d(H;\psi) = a] + \mu(a', H;\hat{\beta})\mathbb{I}[d(H;\psi) = a'].$$

It only requires that either the propensity or mean outcome model to be correctly specified but not both; hence, the *doubly robust* property. In addition to being more robust to model misspecification, AIPW estimators tend to be more efficient than their nonaugmented counterparts (Robins, 2004).

Although its original version was designed for a single-stage treatment regime, it was subsequently adapted to two or more decision points (Zhang et al., 2013; Tao and Wang, 2017; Zhou et al., 2018), where models are

posited for either Q-functions or contrasts.

Outcome weighted learning (OWL) As an alternative direct approach, Zhao et al. (2012) studied the DTR estimation problem as a weighted classification problem–with weights retrospectively determined from clinical outcomes (hence "Outcome Weighted")–and proposed to solve it with ML tools (hence "Learning").

In the case of two treatments, expressed as $A \in \{-1, 1\}$, Qian and Murphy (2011) first showed that the problem can be formulated as a weighted $0-1$ loss in a weighted binary classification problem, where d^* can be estimated as:

$$\hat{d}^* \doteq \arg\max_{d \in \mathcal{D}} \hat{V}^d = \arg\max_{d \in \mathcal{D}} \mathbb{P}_N \left[\frac{\mathbb{I}[A = d(H)]}{\pi(A|H)} Y \right] = \arg\min_{d \in \mathcal{D}} \mathbb{P}_N \left[\frac{\mathbb{I}[A \neq d(H)]}{\pi(A|H)} Y \right].$$

However, due to the discontinuous indicator function, Zhao et al. (2012) proposed to address the optimization problem with a convex surrogate loss function for the $0-1$ loss, corresponding to the *hinge loss* in ML (Hastie et al., 2009). Considering that $d(H)$ can be represented in terms of the sign function $\text{sign}(f(H))$ for some suitable function f, the minimization problem is then expressed as:

$$\hat{f}^* \doteq \arg\min_{f \in \mathcal{F}} \mathbb{P}_N \left[\frac{Y}{\pi(A|H)} \phi(Af(H)) + \lambda_N ||f(H)||^2 \right], \tag{11}$$

where λ_N is a tuning penalty parameter that penalizes the complexity of functions f, $\phi(x) \doteq \max(1-x, 0)$ is the hinge loss, and $||\cdot||$ is the norm function.

An extensive literature has considered some kind of extensions of the standard OWL estimator in Eq. (11), and we outline some of these in Table 2. In particular, Zhao et al. (2015) and Liu et al. (2018) have extended the OWL estimator to a multi-stage setting, proposing the *Backward Outcome Weighted Learning* (BOWL) and *Simultaneous Outcome Weighted Learning* (SOWL) procedures. In the first approach, the stage-t estimator, denoted by $\hat{f}^*_{B,t}$, is obtained recursively as:

$$\hat{f}^*_{B,t} \doteq \arg\min_{f \in \mathcal{F}} \mathbb{P}_N \left[\frac{Y \prod_{\tau=t+1}^{T} \mathbb{I}[A_\tau = \hat{d}^*_\tau(\mathbf{H}_\tau)]}{\prod_{\tau=t}^{T} \pi_\tau(A_\tau|\mathbf{H}_\tau)} \phi(A_t f_t(\mathbf{H}_t)) + \lambda_N ||f_t(\mathbf{H}_t)||^2 \right],$$

where $(\hat{d}^*_{t+1}, \dots \hat{d}^*_T)$ are obtained prior to stage t, and the T-stage estimator does not account for treatments followed afterwards, i.e., $\prod_{\tau=T+1}^{T} \mathbb{I}[A_\tau = \hat{d}^*_\tau(\mathbf{H}_\tau)] \doteq 1$.

The second approach allows simultaneous estimation for all stages. In a two-stage problem, for example, the SOWL estimator, say $\hat{f}^*_S$, is defined as follows:

$$\hat{f}^*_S \doteq \arg\max_{f \in \mathcal{F}} \mathbb{P}_N \left[\frac{Y \psi(A_0 f_0(H_0), A_1 f_1(\mathbf{H_1}))}{\prod_{t=0}^{1} \pi_t(A_t|\mathbf{H_t})} - \lambda_N (||f_0(H_0)||^2 + ||f_1(\mathbf{H_1})||^2) \right],$$

with $\psi(x_1, x_2) \doteq \min(x_1 - 1, x_2 - 1, 0) + 1$ being a concave surrogate for the product of the two (discontinuous) indication functions, introduced to limit computational issues.

Even if numerical examples show that BOWL and SOWL have superior performances compared to existing

direct methods, significant information is lost as t decreases, since only subjects who followed the estimated optimal regime after stage t are used in the backward optimization algorithm. To overcome this problem, an augumented BOWL estimator, leveraging the use of Q-functions, has been proposed and we refer to the original work of Liu et al. (2018) for this extension.

Reference & Acronym	Extension type
Zhao et al. (2015) BOWL + SOWL	**Extension to T-stages**, with $T < \infty$. The authors proposed two methods: one performs an iterative backward OWL (BOWL) estimation, the other a simultaneous OWL (SOWL) estimation.
Liu et al. (2018) AOL	**Extension to negative outcomes and multiple stages**. The authors proposed an augumented version for the weight of the OWL (AOL) integrating OWL and Q-functions. The robust augmentation, making use of predicted pseudo-outcomes from regression models for Q-functions, reduces the variability of weights and improves estimation accuracy.
Zhou et al. (2017) RWL	**Extension to continuous, binary, and count outcomes, and possibility of variable selection**. The authors proposed a general framework, called Residual Weighted Learning (RWL), which employs a *smoothed ramp loss* and derived outcome residuals with a regression model.
Chen et al. (2018) GOWL	**Extension to ordinal treatments and negative outcomes**. The authors proposed a generalized OWL (GOWL) based on a modified loss function and a reformulation of the objective function in the standard OWL.
Zhang et al. (2020) MOML	**Extension to multicategory treatment scenarios and negative outcomes**. The authors used sequential binary methods employing margin-based learning (based on a *large-margin unified machine loss*), which has a special case the standard OWL.
Fu et al. (2019) ROWL	**Extension to outliers, multicategory treatments, and negative outcomes**. The authors proposed a robust OWL (ROWL), based on an angle-based classification structure, designed for multicategory classification problems, and a new family of *robust loss* functions to build more stable DTRs.

Table 2: Extensions of the standard OWL estimator for DTRs.

More recently, under both the direct weighted classification and the indirect blip function frameworks, Luedtke and van der Laan (2016) and colleagues (Montoya et al., 2023) introduced the *SuperLearner* ensemble method (van der Laan et al., 2007) in the DTR arena. Rather than a-priori selecting an estimation framework and algorithm, estimators from both frameworks (and a user-supplied library of candidate algorithms), are combined by using a super-learning based cross-validation selector that seeks to minimize an appropriate cross-validated risk. The full approach is described in Section 3.3 of Montoya et al. (2023).

V-learning for indefinite-horizon problems Most of the work in DTRs focuses on the finite-horizon setting with a very limited literature addressing the in(de)finite case. Yet, for some chronic conditions or those with short or nonfixed time intervals, it may be more natural to assume an in(de)finite time horizon. Tackling this specific setting, Ertefaie and Strawderman (2018) proposed an indirect Q-learning approach, while Luckett et al. (2020) focused on searching an optimal policy over a prespecified class of policies, as in direct methods. Motivated by an mHealth application, where policy estimation is continuously updated in real time as data

accumulate (and starting with small sample sizes), Luckett et al. (2020) introduced *V-learning*. The objective is represented by the value function, written as:

$$V^d(x_t) = \sum_{\tau \geq t} \mathbb{E}\left[\gamma^{\tau-t} Y_{\tau+1} \left(\prod_{v=t}^{\tau} \frac{d(A_v|X_v)}{\pi_v(A_v|S_v)}\right) \middle| X_t = x_t\right], \tag{12}$$

with π an exploration policy, which can be seen as the randomization probability in a randomized trial, and d an arbitrary policy which we want to learn about. Under a time-homogeneous Markov assumption, and provided interchange of the sum and integration is justified, for any function ψ defined on the state space $\mathcal{X}_t$, the value function in Eq. (12) satisfies an importance-weighted variant of the Bellman optimality given by:

$$0 = \mathbb{E}\left[\frac{d(A_t|X_t)}{\pi_t(A_t|S_t)} \left(Y_{t+1} + \gamma V^d(X_{t+1}) - V^d(X_t)\right) \psi(X_t)\right].$$

Let now $V^d(x;\theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^q$, be a model for $V^d(x)$. Assuming that $V^d(x;\theta)$ is differentiable everywhere in θ, for fixed x and d, and denoted with $\psi(x) \doteq \nabla_\theta V^d(x;\theta)$, the proposed estimating equation function is given by:

$$\hat{\Lambda}(\theta) = \mathbb{P}_N\left[\sum_{t=0}^{T} \frac{d(A_t|X_t)}{\pi_t(A_t|S_t)} \left(Y_{t+1} + \gamma V^d(X_{t+1};\theta) - V^d(X_t;\theta)\right) \nabla_\theta V^d(X_t;\theta)\right].$$

An estimate $\hat{\theta}$ can be obtained by minimizing $\hat{M}(\theta) \doteq \hat{\Lambda}(\theta)^T \hat{W}^{-1} \hat{\Lambda}(\theta) + \lambda \mathcal{P}(\theta)$, with $\hat{W}$ a positive definite matrix in $\mathbb{R}^{q \times q}$, λ a tuning parameter and $\mathcal{P} : \mathbb{R}^q \to \mathbb{R}_+$ a penalty function. The optimal estimate $\hat{d}^*$ is then the argmax of $V^d(x;\hat{\theta})$.

3.4 Data sources for constructing DTRs

For the study of DTRs, three main sources of data are typically considered in the literature: i) longitudinal observational studies including cohort studies and EHRs, ii) sequentially randomized studies, with *sequential multiple assignment randomized trial* (SMART) designs (Lavori and Dawson, 2004; Murphy, 2005a) being the "gold standard", and iii) dynamical system models, a tool transposed from control engineering (Rivera et al., 2007). While most real-life studies are based on the first type of data, experimental data sources represent the highest-quality data source for developing DTRs. The third approach has been more peripheral within the DTR literature, typically confined to methodological works aiming at developing and improving existing methodologies. Despite their artificial nature, dynamical systems are built according to biological, behavioral, or social models that simulate realistic individual trajectories. Clinical examples are provided in Thall et al. (2007); Rosenberg et al. (2007), and more behavioral-oriented cases can be found in Rivera et al. (2007); Navarro-Barrientos et al. (2011). For other illustrative examples, we refer to Deliu and Chakraborty (2022). In what follows, the first two types of data sources are discussed.

Longitudinal observational studies Observational data for the study of DTRs include longitudinal trajectories arising from EHRs and other administrative databases or cohort studies (Rosthøj et al., 2006; van der Laan and

Petersen, 2007). They represent a major data source in the biomedical field, constituting an appealing option in scenarios in which a trial would be either cost prohibitive or of concern from an ethical or logistic perspective (e.g., in several chronic diseases such as diabetes or HIV). For this reason, as discussed in Mahar et al. (2021), the estimation of optimal DTRs using observational data has been most concentrated in the area of HIV/AIDS (27, 43%), followed by cancer (8, 13%), and diabetes (6, 10%).

Compared to experimental data arising from randomized clinical trials, using observational data to construct DTRs provides the advantage of evaluating a wider range of treatments at a lower cost. Furthermore, they allow data collection over continuous and indefinite time horizons. However, since the treatments are not randomized within the study, the procedures for drawing causal inference may be affected by the potential presence of time-varying confounders. In particular, the reasons why different individuals receive different treatments or the reasons why one individual receives different treatments at different times are not known with certainty.

Sequentially-randomized studies Although observational data offer a cost-acceptable option and reflect the population's heterogeneity, they present several challenges that make estimation challenging and often subject to various hidden biases. Therefore, randomized data, when available, are preferable for more accurate estimation and stronger statistical inference (Rubin, 1974; Rosenberger et al., 2019). This is especially important when dealing with DTRs since hidden biases can compound over stages. Randomized trials, and most interestingly, SMART designs, are the "gold standard" in DTRs. Randomization coupled with compliance allows causal interpretations to be drawn from statistical association, and are the most effective designs in these multistage medical settings.

A SMART design is characterized by multiple stages of treatment, each stage corresponding to one of the critical decision time point in which randomization occurs. At each subsequent stage, rerandomizations may depend on information collected after previous treatments, but prior to assigning the new treatment, e.g., how well the patient responded to the previous treatment. On the basis of the extent of multiple randomizations, different types of SMARTs can be defined. These include SMARTs in which only nonresponders are rerandomized, and SMARTs in which both responders and nonresponders are rerandomized. In addition, randomization could be made only for one of the initial treatments or for all initial treatments. For a concrete example, see Figure 4 for the schematic of the SMART design adopted for the weight loss management study in Pfammatter et al. (2019). This SMART example involves two stages of treatment and/or experimentation; in general, it may involve as many stages as practically feasible.

3.5 Case study: PROJECT QUIT – FOREVER FREE

Based on a two-stage SMART design, the *PROJECT QUIT – FOREVER FREE* study aimed to develop/compare internet-based (precursors to mobile-based) behavioral interventions for smoking cessation and relapse prevention. The primary aim, based on the six-month-long first stage of the study, i.e., the *PROJECT QUIT*, was to find an optimal multi-factor behavioral intervention to help adult smokers quit smoking; see Strecher et al. (2008) for more details. The second stage, known as *FOREVER FREE*, was a six-month-long follow-on study to help

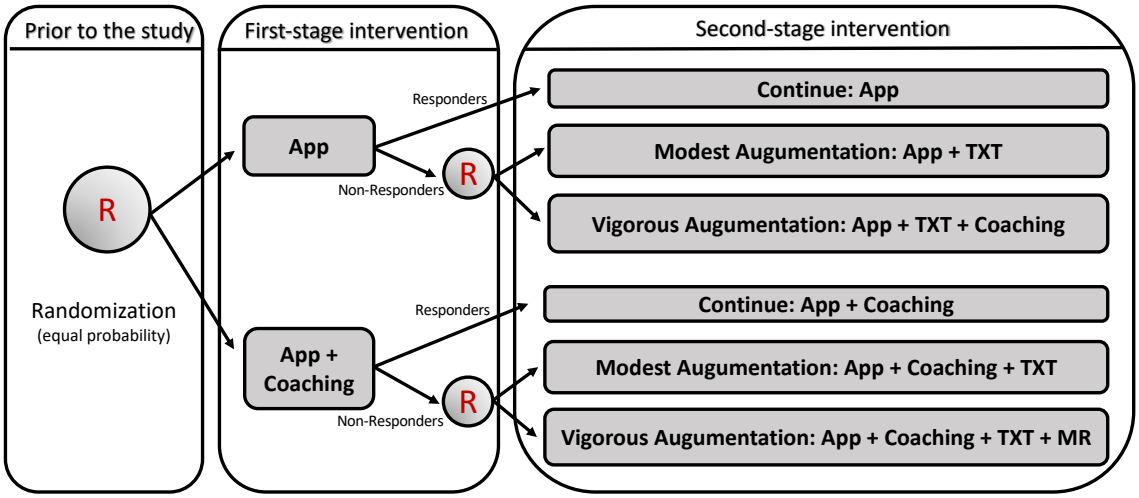

Figure 4: Schematic of the SMART design of the weight loss management study in Pfammatter et al. (2019). App denotes a mobile app, TXT a support text message, and MR meal replacement. The response is defined as a weight loss of at least 0.5 lb on average per week.

PROJECT QUIT participants who quit remain non-smoking, and offer a second chance to those who failed to give up smoking at the previous stage. These two stages were then considered together with the goal of finding an optimal DTR over a twelve-month study period; this was a secondary aim of the main study. RL was not used in the design phase; in other words, this was not an instance of *online learning*. The RL-type learning happened *offline* on completion of data collection, when Q-learning with linear model and a variant (*soft-thresholding*) were employed. The choice of Q-learning was driven by its simplicity and interpretability. Detailed results from this secondary analysis can be found in Chakraborty (2009) and Chakraborty et al. (2010). Here, we only summarize the characteristics of the study in relation to the RL framework and the main challenges we faced.

Selection of interventions and tailoring variables In *PROJECT QUIT*, the original plan was to randomly administer and test six behavioral intervention components (factors), each varied at two levels (highly individually tailored vs. not), according to a 32-cell *fractional factorial design* (FFD). However, due to a program error, one of those factors was not properly implemented. Subsequently, utilizing the factorial structure, the design was "folded" to convert it to a 16-cell FFD and that particular factor was removed from further consideration at the analysis stage. In the primary analysis (Strecher et al., 2008), which was a traditional logistic regression analysis, only two of the five factors were statistically significant. Based on this finding, only these two intervention factors (each at two levels) were considered in the stage-1 Q-learning model. Likewise, various participant-level contextual/tailoring variables were considered in the primary analysis, but only three of them (education, motivation, and self-efficacy) were statistically significant. Again, informed by the primary analysis, only these variables were considered in the stage-1 model of Q-learning, allowing a parsimonious choice of model. In *FOREVER FREE*, originally there were four versions of an active behavioral intervention and a control arm, i.e., five arms in total. But later in the analysis stage, the four versions of the active intervention were found to be minimally different from each other, and hence collapsed. This decision resulted in only two intervention arms at the second stage

of Q-learning.

Reward function The primary outcomes at both stages of the original study were the corresponding seven-day point prevalence of smoking (i.e., whether or not the participant smoked even a single cigarette in the last seven days at six months following the randomization), a dominant measure in the smoking cessation literature. These outcomes were considered as the stage-specific reward functions in Q-learning (Chakraborty et al., 2010). However, the basic operationalization of Q-learning is for continuous outcomes, while the seven-day point prevalence outcomes were binary. Additional Q-learning analysis with a relatively more continuous reward function, the number of months not smoked in the last six months (a secondary outcome in the main study), was also conducted (Chakraborty, 2009). Qualitatively, the results were not too different.

Missing data in the reward variable In *PROJECT QUIT*, 1848 participants were randomly allocated to various interventions, but only 479 of them decided to continue to *FOREVER FREE*; this flexibility was part of the protocol, and hence the remaining *PROJECT QUIT* participants were not considered to be drop-outs for the *FOREVER FREE* part of the study. However, only 1401 out of 1848 stage-1 participants completed the six-month outcome survey; these 1401 participants were treated as complete cases, while the remaining 447 participants were considered drop-outs in stage 1. Similarly, 281 participants (out of 479) who completed the stage-2 six-month survey were treated as complete cases, while the remaining 198 participants were considered drop-outs in stage 2. Descriptive checks revealed that drop-out was more or less uniform across the different intervention arms at both stages. One can employ modern missing data analysis techniques, e.g., *multiple imputation*, before applying Q-learning or other offline RL methods on SMART data to learn about optimal DTRs (see Shortreed et al., 2014, for details). In the case of the *PROJECT QUIT - FOREVER FREE* data, Chakraborty et al. (2010) only presented a complete-case analysis, while Chakraborty (2009) also presented Q-learning analysis of multiply-imputed data.

Statistical inference In DTRs, inference is complicated by the presence of nonregularity (Robins, 2004), a phenomenon characterizing the lack of locally uniform convergence. This can be a result of the sampling distributions of the corresponding estimators changing abruptly as a function of the true underlying parameters. In DTRs, this may occur when two or more treatments produce (nearly) the same mean optimal outcome. To solve this problem, Chakraborty et al. (2010) proposed, first, two alternative ways of shrinking or thresholding values near zero, and then a general method for bootstrapping under nonregularity, i.e., *m-out-of-n bootstrap* (Chakraborty et al., 2013). We refer to these references for the readers interested in this problem.

4 Just-in-time Adaptive Interventions in Digital Health

Digital and mobile health technologies hold the great potential to deliver just-in-time adaptive interventions (JITAIs), i.e., personalized interventions continuously adapted to the real-time contexts of users. As in DTRs,

JITAIs represent a sequence of decision rules tailored to individual users. Their peculiarity is in providing interventions according to the user's in-the-moment context or needs, e.g., time, location, or current activity, including considerations about whether and when the intervention is needed. This feature is enabled by mobile technologies (e.g., wearable devices, accelerometers, or smartphones) that collect data on a continuous basis allowing one to adapt intervention components in real time. In such settings, learning typically occurs online over in(de)finite horizons, and the target policy corresponds to the exploration or behavioral policy. Unlike DTRs, the number of decision points in JITAIs can be hundreds or even thousands, and the intervention can be delivered each minute, hour, or day. This distinctive feature, at least partially, contributed to their increasing popularity in a variety of behavioral domains, including physical activity (Hardeman et al., 2019; Figueroa et al., 2022), addictive disorders in alcohol and drug use (Goldstein et al., 2017; Garnett et al., 2019; Bell et al., 2020), smoking cessation (Naughton, 2017) and obesity/weight management (Aswani et al., 2019). Furthermore, JITAIs have the potential to improve access to quality care in underserved communities and thus alleviate health disparities, a significant public health concern (Liu et al., 2023).

Given the high granularity of the JITAI data, two outcome components play a determinant role in the adaptation and optimization of interventions in real time (Nahum-Shani and Almirall, 2019):

(v) the **proximal outcome(s)** $\{Y_t\}_{t>0}$ directly targeted by an intervention, easily observable and expected to influence a longer-term outcome of interest, according to some mediation theory (MacKinnon et al., 2007);

(vi) the **distal outcome(s)**, representing the long-term health outcome of interest (typically clinical) and ultimate goal of the overall AI.

Taken together with the four components that define a DTR (see Section 3), they represent the six key ingredients of a JITAI. Note that the proximal outcomes can also be used as tailoring variables for guiding later-stage decisions. Unlike DTRs–which target the distal outcome and may or may not have an intermediate (proximal) outcome–in JITAIs, the proximal outcomes represent the direct and in-the-moment target of the intervention. The distal outcome is expected to improve only on the basis of domain knowledge about its relationship with proximal outcomes, but it is not formally included in an optimization problem. A more detailed discussion of the differences between DTRs and JITAIs is provided in Deliu et al. (2024).

4.1 RL methods for JITAIs in digital health

JITAIs are carried out in dynamic environments where the context and needs of individual users can change rapidly (Nahum-Shani et al., 2015, 2018). Therefore, methodologies for delivering JITAIs are required to perform almost continuous learning–with no definite time horizon–and to provide interventions *online* as data accumulate, often utilizing trajectories defined over very short time periods. Note that in such settings, the exploration policy $\boldsymbol{\pi}$ used to collect the samples corresponds to the target policy $\boldsymbol{d}$ we want to improve and optimize; that is, $\boldsymbol{\pi} = \boldsymbol{d}$. However, existing methods for DTRs mainly target a finite-time horizon problem and are implemented offline with backward induction (as in Q-learning); therefore, they are not directly applicable. Furthermore, by carrying over an entire history of an individual, they may not be feasible from a computational perspective. In

such problems, a simplified RL framework, known as the multi-armed bandit (MAB) problem, represents an attractive approach and is increasigly being used within the digital and mobile health domains (Tewari and Murphy, 2017). In what follows, after an illustration of the MAB framework, we describe some popular MAB algorithms that have been used in digital health.

4.1.1 The multi-armed bandit framework

MAB problems can be viewed as a subclass of RL problems (Sutton and Barto, 2018). In the simplest stateless case, the environment does not have any state transitions and actions can be determined according to a single-stage decision-making framework. Generally speaking, the MAB problem (also called the K-armed bandit problem) is a problem in which a limited set of resources (e.g., a group of individuals) must be allocated between competing choices in order to maximize the total expected reward over time. Each of the K choices (i.e., *arms* or actions) provides a different reward, whose probability distribution is specific to that choice. If one knew the expected reward (or value) of each action, then it would be trivial to solve the bandit problem: they would always select the action with the highest value. However, as this information is only partially gained for the selected actions, at each decision time t the agent must trade-off between optimizing its decisions based on acquired knowledge up to time t (*exploitation*) and acquiring new knowledge about the expected rewards of the other actions (*exploration*). The problem conforms to the class of on-policy RL, where the same policy $\boldsymbol{\pi}$ used to explore the actions is evaluated and improved throughout the learning process. For this reason, within this section, we avoid a distinction between $\boldsymbol{\pi}$ and $\boldsymbol{d}$, and we use $\boldsymbol{\pi}$ to refer to both the target policy and the exploration policy.

MAB problems can incorporate some context, which in mapped into appropriate interventions or arms (*contextual* MABs), or solve a context-free task (*non-contextual* MABs), where no side-information is used. In the theory of sequential decision making, contextual MABs occupy a middle ground between non-contextual MABs (Bubeck and Cesa-Bianchi, 2012; Auer et al., 2002b) and full-blown RL.

The most typical assumption is that contexts $\{X_t\}_{t\in\mathbb{N}}$ are independent and identically distributed (i.i.d.) with some fixed but unknown distribution. This means that action A_t at time t has an *in-the-moment* effect on the proximal reward Y_{t+1} at time $t+1$, but not on the distribution of future rewards $\{Y_\tau\}_{\tau\geq t+2}$, for which the i.i.d. property also holds. Under this assumption, one can be completely *myopic* (with $\gamma = 0$ in Eqs. (3) and (8)) and ignore the effect of an action on the distant future in searching for a good policy. In such contextual MABs, and further in the context-free MAB problem, the trajectory distributions are simplified as follows:

$$
\begin{aligned}
P_{\boldsymbol{\pi}} &\doteq p_0(x_0)\prod_{t\geq 0}\pi_t(a_t|x_t)p_{t+1}(x_{t+1},y_{t+1}|x_t,a_t) && \text{[Contextual MAB]}\\
&= p_0(x_0)\prod_{t\geq 0}\pi_t(a_t|x_t)p_{t+1}(x_{t+1})r_{t+1}(y_{t+1}|x_t,x_{t+1},a_t)\\
P_{\boldsymbol{\pi}} &\doteq \prod_{t\geq 0}\pi_t(a_t)r_{t+1}(y_{t+1}|a_t). && \text{[Non-contextual MAB]}
\end{aligned}
$$

As in the general RL problem, the goal of an MAB problem is to select the optimal arm at each time t to

maximize the expected return, alternatively (and with a slightly different nuance) expressed in the bandit literature in terms of minimizing the *total regret*. Formally, denoted by $A_t^* \doteq \arg\max_{a_t \in \mathcal{A}} \mathbb{E}(Y_{t+1}|X_t = x_t, A_t = a_t)$ the optimal arm at time t, we define the *immediate regret* $\Delta(A_t)$ of action A_t as the difference between the expected reward of the optimal arm A_t^* and the expected reward of the ultimately chosen arm A_t, i.e.,

$$\Delta(A_t) \doteq \mathbb{E}(Y_{t+1}|X_t, A_t^*) - \mathbb{E}(Y_{t+1}|X_t, A_t). \tag{13}$$

A nonexhaustive correspondence table between the MAB and JITAI terminologies is reported in **Table 3**.

Table 3: Terminology correspondence between MABs and JITAIs.

Notation	**Terminology**	
	MABs	JITAIs in mHealth
i	Trajectory	User, Subject, Individual
t	Round, Step	Time, Round, Step
X	Context	Context, Contextual Variables
A	Arm	Intervention, Arm
Y	Reward, Payoff	Proximal Outcome
$\mathbf{H}$	Filtration	Filtration
$\boldsymbol{\pi}/\mathbf{d}$	Policy	Just-in-time Adaptive Intervention

With a few exceptions, the contextual MAB algorithms applied in mHealth are based on and rely on the field-specific adaptation of two fundamental contextual bandit approaches: the Upper Confidence Bound (UCB; Li et al., 2010; Chu et al., 2011) and the Thompson sampling (TS; Thompson, 1933; Agrawal and Goyal, 2013) strategies. Exceptions include the Actor-Critic strategy used e.g., in Greenewald et al. (2017).

Contextual bandits with linear UCB Linear Upper Confidence Bound (LinUCB) bandits (Li et al., 2010; Chu et al., 2011) represent an extension of the UCB algorithm (Auer et al., 2002a) for MAB problems to contextual MAB problems. It assumes that the expected reward is a linear function of the context-action feature $f(X_t, A_t) \in \mathbb{R}^{d'}$, i.e., $\mathbb{E}[Y_{t+1}|X_t, A_t] = f(X_t, A_t)^T \mu$, with $\mu \in \mathbb{R}^{d'}$ an unknown reward parameter. In this work, we refer to general features (constructed e.g., via linear basis, polynomials or splines expansion; see e.g., Marsh and Cormier, 2002) rather than a standard linear function that may not capture nonlinearities in the data. Similar modeling considerations have been made in Q-learning (by using, e.g., DNNs).

At each time t, revealed the context X_t, LinUCB calculates the upper confidence bound for the expected reward for all possible actions and then selects the action associated with the highest UCB. Denoted by $U_t(a_t)$ the UCB of arm a_t at time t, Li et al. (2010) and Chu et al. (2011) proposed the formulation:

$$\hat{U}_t(a_t) \doteq \mathbb{E}[Y_{t+1}|X_t, A_t] + \alpha s_t(a_t) = f(X_t = x_t, A_t = a_t)^T \hat{\mu}_t + \alpha s_t(a_t),$$

where $\hat{\mu}_t$ is an estimator of the unknown regression coefficient μ_t and $s_t(a_t)$ is defined as $\sqrt{f(X_t, A_t = a_t)^T B_t^{-1} f(X_t, A_t = a_t)}$, with $B_t \doteq \lambda \mathbb{I}_{d'} + \sum_{\tau=0}^{t-1} f(X_\tau, A_\tau = \tilde{a}_\tau) f(X_\tau, A_\tau = \tilde{a}_\tau)^T$. B_t is computed recursively at each time step t, by taking into account the context-action features associated with the optimal actions $\{\tilde{a}_\tau \doteq \arg\max_{a_\tau \in \mathcal{A}} U_\tau(a_\tau)\}_{\tau=0,1,\ldots,t-1}$ estimated at previous rounds. Note that the first part

$f(X_t, A_t = a_t)^T \hat{\mu}_t$ reflects the current estimate of the reward, while the second part $s_t(a_t)$ is an indication of its uncertainty; thus, it naturally balances between exploration and exploitation. The tuning parameter $\alpha > 0$ balances the trade-off between exploration and exploitation: small values of α favor exploitation, while larger values of α favor exploration.

Moving from pure bandit and statistical theory to real-world digital health applications, the use of LinUCB has been reported in Forman et al. (2019), among others. Here, in the context of behavioral weight loss and maintenance, a pilot study has been conducted to evaluate the feasibility and acceptability of an RL-based intervention. Participants were randomized into a nonoptimized, an individually optimized (individual reward maximization), and a group optimized (group reward maximization) group. The study showed the advantages of the RL-based optimized groups in terms of the outcome of interest, not only being feasible to deploy and acceptable to participants and coaches, but also achieving desirable results at roughly one-third the cost.

Contextual bandits with linear Thompson sampling Under the same linear assumption of LinUCB, Agrawal and Goyal (2013) proposed a randomized version of the latter, based on a generalization of the TS technique for i.i.d. contextual MAB problems. Based on the Bayesian framework, the idea of TS is to randomly allocate each arm according to its posterior probability of being optimal. More specifically, assuming a Gaussian prior for the μ parameter $\mu \sim \mathcal{N}(\mathbf{0}_{d'}, \nu^2 \mathbb{I}_{d'})$ and a Gaussian distribution for the reward $Y_t|\mu, f(X_t, A_t) \sim \mathcal{N}(f(X_t, A_t)^T \mu, \nu^2)$, for some $\nu > 0$, at each time t the optimal arm $\tilde{a}_t$ is the one that maximizes the 'a-posteriori' estimated expected reward, i.e., $f(X_t, A_t)^T \tilde{\mu}_t$. The posterior nature is reflected in $\tilde{\mu}_t$, which represents a sample from the posterior distribution, computed recursively and given by $\mathcal{N}(\hat{\mu}_t, \nu^2 B_t^{-1})$, with $\hat{\mu}_t \doteq B_t^{-1} b_t$, where $B_t \doteq \mathbb{I}_{d'} + \sum_{\tau=0}^{t-1} f(X_\tau, A_\tau = \tilde{a}_\tau) f(X_\tau, A_\tau = \tilde{a}_\tau)^T$ and $b_t \doteq \sum_{\tau=0}^{t-1} f(X_\tau, A_\tau = \tilde{a}_\tau) Y_{\tau+1}(X_\tau, A_\tau = \tilde{a}_\tau)$. The policy π at each time t is thus explicitly defined as:

$$\pi_t(a) = \mathbb{P}\Big(\mathbb{E}[Y_{t+1} \mid X_t = x_t, A_t = a] \geq \mathbb{E}[Y_{t+1} \mid X_t = x_t, A_t = a'], \forall a' \neq a \mid \mathcal{F}_{t-1}\Big),$$

where the conditioning term $\mathcal{F}_{t-1}$ reflects the posterior nature of this strategy.

Given all the trajectory information up to time t, $\mathcal{T}_{t-1} = \{(X_\tau, A_\tau, Y_{\tau+1})\}_{\tau=0,1,\ldots,t-1}$ and $f(X_t, A_t)$, LinUCB is deterministic and allows exploration through the uncertainty term $\alpha s_t(a_t)$. On the other hand, in TS, exploration is given by the random draws from the posterior distribution. Note that the standard deviation $\alpha s_t(a_t)$ characterizing LinUCB has the same order as the standard deviation of the posterior distribution of the reward $Y_t|\mu_t, f(X_t, A_t) \sim \mathcal{N}(f(X_t, A_t = a_t)^T \hat{\mu}_t, \nu^2 f(X_t, A_t = a_t)^T B_t^{-1} f(X_t, A_t = a_t))$ used in TS, where $f(X_t, A_t = a_t)^T B_t^{-1} f(X_t, A_t = a_t) = s_t(a_t)$ by definition.

Actor-critic contextual bandits Specifically addressing personalized mHealth intervention problems, Lei (2016) proposed to use a particular RL setting, called actor-critic (Grondman et al., 2012), based on which both policies and value functions are learned. The "actor" is the component that learns policies, and the "critic" is the component that learns about whatever policy the actor is currently following to "criticize" its choices (Sutton and Barto, 2018).

Assuming a linear model for the reward $Y_{t+1} = f(X_t, A_t)^T \mu_t + \epsilon_{t+1}$, with $f(X_t, A_t) \in \mathbb{R}^{d'}$, $\mu \in \mathbb{R}^{d'}$ and i.i.d. error terms $\{\epsilon_t\}_{t\in\mathbb{N}}$ with mean 0 and variance σ^2, and taking into account a binary action space $\mathcal{A} = \{0, 1\}$, Lei (2016) formulated an online policy learning procedure as a contextual bandit problem, and proposed a class of parametrized stochastic policies with $\mathbb{P}(A = 1|X = x) = \pi(1|x; \theta) = \frac{e^{g(x)^T\theta}}{1+e^{g(x)^T\theta}}$, and $g(x)$ a p-dimensional policy feature.

To maintain variety of treatment and increasing engagement, a stochastic chance constraint related to a parametrized stationary policy is introduced. In this specific case, if we denote with $\{\pi(a|x; \theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$ the class of parameterized stochastic policies, the stochasticity constraint has the form:

$$\mathbb{P}(\pi_{min} \leq \pi(A = 1|X; \theta) \leq 1 - \pi_{min}) \geq 1 - \alpha, \tag{14}$$

with $\pi_{min} \in (0, .5)$ and $\alpha \in (0, 1)$ controlling the amount of stochasticity.

An optimal policy can then be obtained by maximizing the expected reward under the policy $\pi(a|x; \theta)$, i.e., $V^\pi(\theta) \doteq \mathbb{E}_{\pi_\theta}(Y)$, subject to the constraint in Eq. (14). Solving this constrained optimization problem involves a major difficulty given the nonconvex constraint on θ, which involves also some nonsmoothness. To circumvent this difficulty, the authors relax the above constraint by bounding the probability in Eq. (14) using Markov's inequality, and then solving the constrained optimization problem using the Lagrangian function $J_\lambda(\theta)$, with λ the Lagrangian multiplier. That is, for a fixed λ, the optimal policy $\pi^* \doteq \pi_{\theta^*}$ is the one with θ^* given by:

$$\theta^* \doteq \arg\max_{\theta\in\Theta} J_\lambda(\theta),$$

where $J_\lambda(\theta)$, also referred to as regularized average reward, is defined as

$$\begin{aligned} J_\lambda(\theta) &\doteq V^\pi(\theta) - \lambda\theta^T\mathbb{E}(g(X)g(X)^T)\theta \\ &= \mathbb{E}_{\pi_\theta}(Y) - \lambda\theta^T\mathbb{E}(g(X)g(X)^T)\theta \\ &= \mathbb{E}_{p(x)}\mathbb{E}_{\pi(a|x;\theta)}[E(Y|X = x; A = a)] - \lambda\theta^T\mathbb{E}(g(X)g(X)^T)\theta. \end{aligned}$$

Being $J_\lambda(\theta)$ unknown, its MC version is considered:

$$\widehat{J}_\lambda(\theta) = \mathbb{P}_N\left[\sum_a E(Y|X = x; A = a)\pi(a|X = x; \theta) - \lambda\theta^T(g(x)g(x)^T)\theta\right],$$

where $\mathbb{P}_N$ denotes the empirical average on N i.i.d. samples.

For estimating the expected reward $E(Y|X = x; A = a)$, the linear assumption $\mathbb{E}(Y_{t+1}|X_t = x, A_t = a) = f(x, a)^T\mu$ is used, and the L_2 norm penalized least square estimator is considered. This helps to overcome the full rank requirement of the matrix $\sum_{t=0}^T f(X_t, A_t)f(X_t, A_t)^T$ at the beginning of the experiment when running the algorithm online.

While the proposed algorithm is formulated for a binary action space, we notice that this class of methods has been originally introduced to overcome the limitation of methods that fail to address complex action space

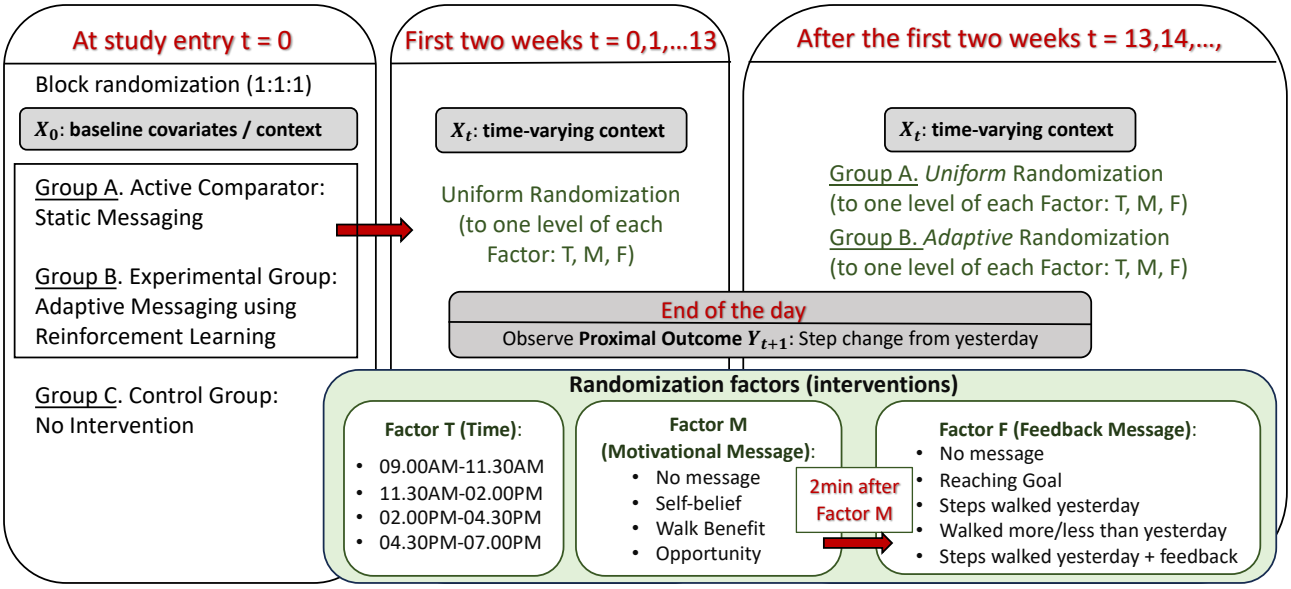

Figure 5: Schematic of the MRT design of the *DIAMANTE* Study.

problems (e.g., tabular Q-learning). Thus, the greater advantage of an actor-critic approach occurs in settings with action spaces more complex than the binary case, which may be solved by simpler methods such as LinTS or LinUCB.

4.2 Data sources for building JITAIs in digital health

Typical experimental designs for building JITAIs are represented by *factorial experiments* (Collins et al., 2009), or, most notably, *micro-randomized trials* (MRTs; Klasnja et al., 2015). In MRTs, individuals are randomized hundreds or thousands of times over the course of the study, and, in a typical multicomponent intervention study, the multiple components can be randomized concurrently, making micro-randomization a form of a sequential factorial design. The goal of these trials is to optimize mHealth interventions while assessing the causal effects of each randomized intervention component and evaluating whether the intervention effects vary with time or the current context of individuals. A review of this cutting-edge design, covering both its classical variants and its adaptive counterpart, together with the associated statistical challenges, is presented in Liu et al. (2023).

As an illustrative example, to better understand the characteristics and value of an MRT, let us now consider the *DIAMANTE* study design of a physical activity trial in Figure 5. In this study, the intervention options (i.e., each unique combination of the factor levels) include whether or not to send a text message, which type of message to deliver, and at which time; the proximal outcome is the change in the number of steps the person walked today from yesterday; and the context is given by a set of user's individual variables such as baseline health status or study day.

At the macro level, to assess the benefits of optimized JITAIs, in addition to evaluating the causal role of each intervention, users are randomized to different study groups (see Figure 5), including a static (control) group, a uniform random (nonoptimized) group, and an RL-based (adaptive, optimized) group. In noncontrol groups, users are randomized every day to receive a combination of the different factors' levels, delivered within different

time frames. The adaptive RL-based optimized group strategy is illustrated in more detail in Section 4.3.

4.3 Case study: DIAMANTE trial

The *DIAMANTE* trial (Aguilera et al., 2020) is an mHealth study that we designed in collaboration with a team of clinicians, phychologists, and computer scientists, among others. The general objective of the study was to encourage users to become more physically active by sending them suitable text messages, and the design involved RL architectures. The overview of the trial is given in Figure 5, and preliminary (pilot data) results are presented in Figueroa et al. (2022).

During the pilot study, we implemented an adaptive RL mechanisms on a daily basis for deciding on: 1) the feedback message, 2) the motivation message, and 3) the timing of the message. Each combination of levels of the three experimental factors represented an arm or action. To increase personalization, the decision about which message to send also took into consideration contextual variables: time-independent variables such as baseline sociodemographic information and time-dependent covariates, including the day of the week (Monday-Sunday), the data steps of the previous day and the number of days since messages from different categories were sent. For this study, we adopted a MAB strategy adapted to the specific setting of mHealth and the high dimensionality of the context, which we detail below.

Algorithm In this trial, we employed RL only for one of the groups, regarded as the adaptive experimental group. We proposed the contextual linear TS algorithm and decided to implement it after the first two weeks, during which text messages were sent uniformly at random (analogous to an initial "burn-in" period, or, more appropriately, an "internal pilot" to acquire some prior data to feed into the main algorithm). The choice of TS was motivated by several reasons. First, its empirical and theoretical properties have been well studied and have shown great theoretical and empirical performances (Chapelle and Li, 2011; Agrawal and Goyal, 2013). Second, it is computationally efficient, and thus particularly suitable for online learning (Russo et al., 2018). Third, it is a randomized algorithm and, as such, mitigates different forms of biases and allows causal inference (Rosenberger et al., 2019). Finally, TS has been widely applied in real-world applications, including mHealth (Liao et al., 2020) and showed promising results, even with small amounts of data (Agrawal and Goyal, 2013).

Regularization We used a Bayesian linear regression setting with a *Normal-Inverse-Gamma* (NIG) prior for the regression coefficients (mean and variance). This choice keeps the coefficients small and minimizes overfitting, providing some form of regularization. A Gaussian distribution is assumed for the reward, i.e.,

$$Y_t | f(X_t, A_t), \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(f(X_t, A_t)^T \boldsymbol{\beta}, \sigma^2), \quad t = 1, 2, \ldots$$

with unknown coefficient vector $\boldsymbol{\beta}$ and unknown variance $\sigma^2 > 0$. Following a Bayesian framework, the

unknown parameters are assumed to be jointly distributed as a multivariate NIG, that is,

$$(\boldsymbol{\beta}, \sigma^2)|\boldsymbol{\mu}_\beta, \Sigma_\beta, a, b \sim \text{NIG}_{d+1}(\boldsymbol{\mu}_\beta, \Sigma_\beta, a, b),$$

where $\boldsymbol{\mu}_\beta \in \mathbb{R}^{d+1}$, with a, $b \in \mathbb{R}_{>0}$, are fixed and known prior hyper-parameters. When Σ_β, the covariance structure between the $\boldsymbol{\beta}$ vector, is known, the joint prior distribution can be simplified through a hierarchical representation as:

$$\begin{aligned}\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \Sigma_\beta, \sigma^2 &\sim \mathcal{N}_{d+1}(\boldsymbol{\mu}_\beta, \sigma^2\Sigma_\beta),\\ \sigma^2|a, b &\sim \text{IG}(a, b).\end{aligned}$$

Note that this also simplifies the sampling process of the TS algorithm. In fact, to perform a posterior sampling from the NIG posterior distribution, it is enough to sample the unknown variance and mean parameters from their updated Inverse-Gamma and Normal posteriors, which we denote with $\text{IG}(a^*, b^*)$ and $\mathcal{N}_{d+1}(\boldsymbol{\mu}^*, \tilde{\sigma}^2_{(t)}\Sigma^*)$, respectively, where

$$\begin{aligned}\boldsymbol{\mu^*} &= \left(\Sigma_\beta^{-1} + f(X_t, A_t)^T f(X_t, A_t)\right)^{-1}\left(\Sigma_\beta^{-1}\boldsymbol{\mu}_\beta + f(X_t, A_t)^T Y_t\right),\\ \Sigma^* &= \left(\Sigma_\beta^{-1} + f(X_t, A_t)^T f(X_t, A_t)\right)^{-1},\\ a^* &= a + \frac{n}{2},\\ b^* &= b + \frac{1}{2}\left(\boldsymbol{\mu}_\beta^T\Sigma_\beta^{-1}\boldsymbol{\mu}_\beta + Y^TY - \boldsymbol{\mu^*}^T\Sigma^{*-1}\boldsymbol{\mu^*}\right),\\ \tilde{\sigma}^2_{(t)} &\sim \text{IG}(a^*, b^*).\end{aligned}$$

The resulting vector $\left\{\tilde{\boldsymbol{\beta}}_{(t)}, \tilde{\sigma}^2_{(t)}\right\}_{t=1}^T$, with $\tilde{\sigma}^2_{(t)} \sim \text{IG}(a^*, b^*)$ and $\tilde{\boldsymbol{\beta}}_{(t)} \sim \mathcal{N}_{d+1}(\boldsymbol{\mu}^*, \tilde{\sigma}^2_{(t)}\Sigma^*)$ provides samples from the joint NIG posterior, while $\left\{\tilde{\boldsymbol{\beta}}_{(t)}\right\}_{t=1}^T$ and $\left\{\tilde{\sigma}^2_{(t)}\right\}_{t=1}^T$ provide samples from the marginal Normal and IG posterior, respectively. Based on the posterior samples, at each iteration t, the optimal arm $\tilde{a}_{(t)}$ is the one that maximizes the posterior estimated expected reward, $f(X_t, A_t)^T\tilde{\boldsymbol{\beta}}_{(t)}$, where the posterior nature is reflected in $\tilde{\boldsymbol{\beta}}_{(t)}$.

Habituation In many real-world scenarios, temporal changes in the reward distribution are an intrinsic characteristic of the problem, and the stationary assumption may be too simplistic. For example, one may expect that the effectiveness of a specific intervention would deteriorate over time or with continued exposure to that intervention. In mHealth, or more generally in behavioral sciences, this phenomenon, known as *habituation* (Epstein et al., 2009), is a recognized pattern. In a text-messaging application such as the DIAMANTE study, sending the same message category repeatedly over time can eventually result in a decreased response or, even worse, disengagement with the program. Consequently, an arm that was optimal for an individual for a certain number of initial days might lose its effectiveness, and the algorithm should quickly adapt to this shift. To account for this departure from the stationarity assumption, we

followed a *recovery bandit* (Pike-Burke and Grünewälder, 2019) approach and modeled the outcome of interest (daily step change) as a function of the number of times since each given text-message category was last sent. Consider K different arms (e.g., message categories) and a fixed number of days T in which only one category can be sent. Now, for each text-message category $A_j, j = 1, \dots K$ and day $t = 0, \dots, T$, we denote by $Z_{A_j,t}$ the number of days since category A_j was last played, where $Z_{A_j,t} \in \mathcal{Z} = \{0, \dots, Z_{\max}\}$ for a finite $Z_{\max} \in \mathbb{N}$. More specifically, at day $t+1$, we have that for each $A_j, j = 1, \dots K$,

$$Z_{A_j,t+1} = \begin{cases} 0 & \text{if } A_t = A_j \text{ (or } A_{j,t} = 1), \\ \min\{Z_{\max}, Z_{A_j,t} + 1\} & \text{if } A_t \neq A_j \text{ (or } A_{j,t} = 0). \end{cases}$$

The variable $A_{j,t}$ is the dummy version of the message category representing whether the category A_j was chosen at time t: $A_{j,t} = 1$ means that on day t category A_j was assigned ($A_t = A_j$), while $A_{j,t} = 0$ means that on day t a category different from A_j was selected ($A_t \neq A_j$).

Let now $\overline{\mathbf{Z}}_t \doteq (\overline{Z}_{A_1,t}, \dots, \overline{Z}_{A_K,t}) \doteq (Z_{\max} - Z_{A_1,t}, \dots, Z_{\max} - Z_{A_K,t})$, for $t = 1, \dots, T$, be the vector of these auxiliary variables, which we name habituation or recovery context. The idea is that, based on the closeness in time of a certain text message category, the reward might be positively or negatively affected. Particularly, with (negative) habituation, we refer to the case where sending the same category consecutively may cause habituation and loss of its potential effect in terms of step change. In this setting, a higher $\overline{Z}_{A_j,t}$ represents a higher degree of habituation at time t related to the category A_j, indicating a higher loss in terms of reward if the same message category is sent over time. In line with the behavioral science literature, we hypothesize that in a fixed number of rounds $Z_{\max}$, a specific category will be exempted from habituation if not sent. Based on this reasoning, at time t, we model the daily step change of category A_j as a linear function of this arm, of the time and/or action invariant context $\mathbf{X}_t$ and of the related arm dependent contextual variable $\overline{Z}_{A_j,t}$, i.e.,

$$\mathbb{E}[Y_t | \mathbf{X}_t, A_t = A_j, \overline{Z}_{A_j,t}] = f(\mathbf{X}_t, A_j, \overline{Z}_{A_j,t})^T \boldsymbol{\beta}, \quad j = 1, \dots, K.$$

Thus, the expected reward of every arm changes at each round t, and this change depends on whether arm A_j was previously played and how many rounds ago.

Reward variable As in DTRs, choosing the appropriate reward variable is a major issue, and in mHealth it strongly depends on the mobile or wearable instrument. In this study, we used the daily step counts (collected by the pedometer on the participants' personal phones) as the proximal outcome for physical activity. To account for users' baseline walking propensity, we decided to consider the step change from one day to another, which also showed a closer Gaussian shape.

Missing data in the reward variable To date, there has been a lack of mHealth studies addressing this problem. However, in an online experimental setting, it is particularly relevant, as it can impact subsequent

selection of interventions when reward is missing. In this work, we set as missing all the 0 step counts; this is done in line with the existing literature, which suggests that this outcome is due to technical errors of the device, or simply users' forgetfulness to carry their phones when walking. Then, taking an exploratory approach, we performed multiple imputation of missing data during post-data collection analysis, more specifically as a sensitivity analysis. During data collection, we used the last observation carried forward technique.

5 Final Remarks

The content of this work summarizes and highlights the increasing potential and interest in RL and AI in general, to improve healthcare and promote healthy behaviors. However, despite remarkable theoretical results and aside from a few examples, fulfilling the vision of precision health and well-being for all is still far from realization. Although the opportunity and potential of AI to make health systems more efficient, sustainable, and equitable is visible, its real-world application to create personalized and effective care remains largely untapped.

In fact, despite remarkable theoretical results, only a few studies applied RL in real life. Moreover, in many cases, applications simply used RL approaches for solving the problem in relatively simplified settings, thus exhibiting a number of shortcomings and practical limitations and posing interesting technical challenges and open problems. The following are a list of nonexhaustive questions that still need to be fully addressed. How can we better understand and interpret the process of an RL algorithm, which often acts in a black box expressed by, for instance, deep neural networks? How can we adequately adapt the RL strategy to complex disease and behavioral scenarios? This requires formalizing appropriate relationships in the RL process, particularly for the reward function, taking into account domain knowledge about each specific setting, its multiple objectives, and the potential presence of nonstationarities or unstructured data. While, for instance, several software packages exist for implementing many of the reviewed algorithms, these are often suitable only under specific cases (e.g., only continuous and positive rewards) and require detailed knowledge about the software.

In addition, more work remains to be done to test, validate, and change healthcare practices, navigating complex ethical, technical, and human-centered challenges in privacy, transparency, and fairness (Chien et al., 2022), among others (see also Deliu et al., 2024). Finally, precision health is a highly multidisciplinary domain–intersecting behavioral and clinical domains, as well as methodological area–and is still in its infancy, having been formalized as a progeny of precision medicine only in the last decade (Ryan et al., 2021). The challenge is thus to unravel the potential of AI in precision health in a cohesive and synergistic way, taking into account the methodological details and their practical adoption. This is the approach we followed in this work, providing a detailed survey of the methodological framework of interest and discussing its benefits in the context of precision and digital health with concrete case studies.

References

Aaltonen, L. A., Abascal, F., Abeshouse, A., and others, and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93.

Agrawal, R. and Prabakaran, S. (2020). Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity*, 124(4):525–534.

Agrawal, S. and Goyal, N. (2013). Thompson Sampling for Contextual Bandits with Linear Payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28(3), pages 127–135. PMLR.

Aguilera, A., Figueroa, C. A., Hernandez-Ramos, R., Sarkar, U., Cemballi, A., Gomez-Pathak, L., Miramontes, J., Yom-Tov, E., Chakraborty, B., Yan, X., Xu, J., Modiri, A., Aggarwal, J., Williams, J. J., and Lyles, C. R. (2020). mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE Study. *BMJ Open*, 10(8):e034723.

All of Us Research Program Investigators, Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., and Dishman, E. (2019). The "All of Us" Research Program. *The New England Journal of Medicine*, 381(7):668–676.

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., Sprosen, T., and Collins, R. (2012). UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1(3):123–126.

Almirall, D., Nahum-Shani, I., Sherwood, N. E., and Murphy, S. A. (2014). Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Translational Behavioral Medicine*, 4(3):260–274.

Aswani, A., Kaminsky, P., Mintz, Y., Flowers, E., and Fukuoka, Y. (2019). Behavioral Modeling in Weight Loss Interventions. *European Journal of Operational Research*, 272(3):1058–1072.

Atan, O., Jordon, J., and Van Der Schaar, M. (2018). Deep-Treat: Learning Optimal Personalized Treatments From Observational Data Using Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77.

Beaglehole, R., Bonita, R., Horton, R., Adams, C., Alleyne, G., Asaria, P., Baugh, V., Bekedam, H., Billo, N., Casswell, S., Cecchini, M., Colagiuri, R., Colagiuri, S., Collins, T., Ebrahim, S., Engelgau, M., Galea, G., Gaziano, T., Geneau, R., Haines, A., Hospedales, J., Jha, P., Keeling, A., Leeder, S., Lincoln, P., McKee, M.,

Mackay, J., Magnusson, R., Moodie, R., Mwatsama, M., Nishtar, S., Norrving, B., Patterson, D., Piot, P., Ralston, J., Rani, M., Reddy, K. S., Sassi, F., Sheron, N., Stuckler, D., Suh, I., Torode, J., Varghese, C., and Watt, J. (2011). Priority actions for the non-communicable disease crisis. *The Lancet*, 377(9775):1438–1447.

Bell, L., Garnett, C., Qian, T., Perski, O., Potts, H. W. W., and Williamson, E. (2020). Notifications to Improve Engagement With an Alcohol Reduction App: Protocol for a Micro-Randomized Trial. *JMIR research protocols*, 9(8):e18690.

Bellman, R. (1957). *Dynamic programming*. Dover Publications, Mineola, N.Y.

Bertsekas, D. P. (2019). *Reinforcement Learning and Optimal Control*. Athena Scientific, Belmont, Massachusetts, 2 edition.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information science and statistics. Springer, New York.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. *Cell*, 173(7):1581–1592.

Chakraborty, B. (2009). *A Study of Non-regularity in Dynamic Treatment Regimes and Some Design Considerations for Multicomponent Interventions*. PhD Thesis, University of Michigan.

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for Optimal Dynamic Treatment Regimes Using an Adaptive m-out-of-n Bootstrap Scheme. *Biometrics*, 69(3):714–723.

Chakraborty, B. and Moodie, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Statistics for Biology and Health. Springer, New York, NY.

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic Treatment Regimes. *Annual Review of Statistics and Its Application*, 1(1):447–464.

Chapelle, O. and Li, L. (2011). An Empirical Evaluation of Thompson Sampling. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, volume 24, pages 2249–2257.

Chen, G., Zeng, D., and Kosorok, M. R. (2016). Personalized Dose Finding Using Outcome Weighted Learning. *Journal of the American Statistical Association*, 111(516):1509–1521.

Chen, J., Fu, H., He, X., Kosorok, M. R., and Liu, Y. (2018). Estimating individualized treatment rules for ordinal treatments. *Biometrics*, 74(3):924–933.

Chien, I., Deliu, N., Turner, R., Weller, A., Villar, S., and Kilbertus, N. (2022). Multi-disciplinary fairness considerations in machine learning for clinical trials. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 906–924, Seoul Republic of Korea. ACM.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual Bandits with Linear Payoff Functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Collins, F. S. and Varmus, H. (2015). A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9):793–795.

Collins, L. M., Chakraborty, B., Murphy, S. A., and Strecher, V. (2009). Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. *Clinical Trials*, 6(1):5–15.

Collins, L. M., Murphy, S. A., and Bierman, K. L. (2004). A Conceptual Framework for Adaptive Preventive Interventions. *Prevention Science*, 5(3):185–196.

De Lara, M., De Lara, M., and Doyen, L. (2008). *Sustainable management of natural resources: mathematical models and methods*. Environmental science and engineering Subseries Environmental science. Springer, Berlin Heidelberg.

Deliu, N. and Chakraborty, B. (2022). Dynamic Treatment Regimes for Optimizing Healthcare. In Chen, X., Jasin, S., and Shi, C., editors, *The Elements of Joint Learning and Optimization in Operations Management*, Springer Series in Supply Chain Management, pages 391–444. Springer International Publishing, Cham.

Deliu, N., Williams, J. J., and Chakraborty, B. (2024). Reinforcement Learning in Modern Biostatistics: Constructing Optimal Adaptive Interventions. *International Statistical Review*, In Press.

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, 132(20):1920–1930.

Epstein, L. H., Temple, J. L., Roemmich, J. N., and Bouton, M. E. (2009). Habituation as a determinant of human food intake. *Psychological Review*, 116(2):384–407.

Ertefaie, A. and Strawderman, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4):963–977.

Figueroa, C. A., Deliu, N., Chakraborty, B., Modiri, A., Xu, J., Aggarwal, J., Jay Williams, J., Lyles, C., and Aguilera, A. (2022). Daily Motivational Text Messages to Promote Physical Activity in University Students: Results From a Microrandomized Trial. *Annals of Behavioral Medicine*, 56(2):212–218.

Forman, E. M., Kerrigan, S. G., Butryn, M. L., Juarascio, A. S., Manasse, S. M., Ontañón, S., Dallal, D. H., Crochiere, R. J., and Moskow, D. (2019). Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of Behavioral Medicine*, 42(2):276–290.

Fu, S., He, Q., Zhang, S., and Liu, Y. (2019). Robust outcome weighted learning for optimal individualized treatment rules. *Journal of Biopharmaceutical Statistics*, 29(4):606–624.

Gambhir, S. S., Ge, T. J., Vermesh, O., and Spitler, R. (2018). Toward achieving precision health. *Science Translational Medicine*, 10(430):eaao3612.

Garnett, C., Crane, D., West, R., Brown, J., and Michie, S. (2019). The development of *Drink Less* : an alcohol reduction smartphone app for excessive drinkers. *Translational Behavioral Medicine*, 9(2):296–307.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. *The Annals of Statistics*, 40(1).

Goldstein, S. P., Evans, B. C., Flack, D., Juarascio, A., Manasse, S., Zhang, F., and Forman, E. M. (2017). Return of the JITAI: Applying a Just-in-Time Adaptive Intervention Framework to the Development of m-Health Solutions for Addictive Behaviors. *International Journal of Behavioral Medicine*, 24(5):673–682.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18.

Greenewald, K., Tewari, A., Klasnja, P., and Murphy, S. (2017). Action centered contextual bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 5979–5987, Red Hook, NY, USA. Curran Associates Inc.

Grondman, I., Busoniu, L., Lopes, G. A. D., and Babuska, R. (2012). A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1291–1307.

Guan, Y., Li, S. E., Duan, J., Li, J., Ren, Y., Sun, Q., and Cheng, B. (2021). Direct and indirect reinforcement learning. *International Journal of Intelligent Systems*, 36(8):4439–4467.

Hardeman, W., Houghton, J., Lane, K., Jones, A., and Naughton, F. (2019). A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 16(1):31.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Hernan, M. A. and Robins, J. M. (2023). *Causal Inference: What If*. CRC Press, Boca Raton.

Istepanian, R. S. H., Laxminarayan, S., and Pattichis, C. S., editors (2006). *M-Health: Emerging Mobile Health Systems*. Topics in Biomedical Engineering. International Book Series (ITBE). Springer, New York, N.Y.

Johnson, K. B., Wei, W., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., and Snowdon, J. L. (2021). Precision Medicine, AI, and the Future of Personalized Health Care. *Clinical and Translational Science*, 14(1):86–93.

Kitagawa, T. and Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2):591–616.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(Suppl):1220–1228.

Kosorok, M. R. and Laber, E. B. (2019). Precision Medicine. *Annual Review of Statistics and Its Application*, 6(1):263–286.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17.

Laber, E. B., Lizotte, D. J., and Ferguson, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70(1):53–61.

Laber, E. B. and Zhao, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514.

Lavori, P. W. and Dawson, R. (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38.

Lavori, P. W. and Dawson, R. (2004). Dynamic treatment regimes: practical design considerations. *Clinical Trials*, 1(1):9–20.

Lei, H. (2016). *An Online Actor Critic Algorithm and a Statistical Decision Procedure for Personalizing Intervention*. PhD Thesis, University of Michigan.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, Raleigh North Carolina USA. ACM.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020). Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22.

Liu, X., Deliu, N., and Chakraborty, B. (2023). Microrandomized Trials: Developing Just-in-Time Adaptive Interventions for Better Public Health. *American Journal of Public Health*, 113(1):60–69.

Liu, Y., Logan, B., Liu, N., Xu, Z., Tang, J., and Wang, Y. (2017). Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 380–385, Park City, UT. IEEE.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens: Augmented Outcome-weighted Learning. *Statistics in Medicine*, 37(26):3776–3788.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020). Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning. *Journal of the American Statistical Association*, 115(530):692–706.

Luedtke, A. R. and van der Laan, M. J. (2016). Super-Learning of an Optimal Dynamic Treatment Rule. *The international journal of biostatistics*, 12(1):305–332.

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation Analysis. *Annual Review of Psychology*, 58(1):593–614.

Mahar, R. K., McGuinness, M. B., Chakraborty, B., Carlin, J. B., IJzerman, M. J., and Simpson, J. A. (2021). A scoping review of studies using observational data to optimise dynamic treatment regimens. *BMC Medical Research Methodology*, 21(1):39.

Marsh, L. and Cormier, D. (2002). *Spline Regression Models*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J., and Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Montoya, L. M., van der Laan, M. J., Luedtke, A. R., Skeem, J. L., Coyle, J. R., and Petersen, M. L. (2023). The optimal dynamic treatment rule superlearner: considerations, performance, and application to criminal justice interventions. *The International Journal of Biostatistics*, 19(1):217–238.

Moodie, E. E. M., Dean, N., and Sun, Y. R. (2014). Q-Learning: Flexible Learning About Useful Utilities. *Statistics in Biosciences*, 6(2):223–243.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481.

Murphy, S. A. (2005b). A Generalization Error for Q-Learning. *The Journal of Machine Learning Research*, 6:1073–1097.

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M. B., Neil, D., Saffari, A., Mead, R., Hautbergue, G. M., Holbrook, J. D., and Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews. Neurology*, 16(8):440–456.

Nahum-Shani, I. and Almirall, D. (2019). An Introduction to Adaptive Interventions and SMART Designs in Education (NCSER 2020-001). Technical report, U.S. Department of Education. Washington, DC: National Center for Special Education Research.

Nahum-Shani, I., Hekler, E. B., and Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology*, 34(Suppl):1209–1219.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. (2018). Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine*, 52(6):446–462.

Naughton, F. (2017). Delivering "Just-In-Time" Smoking Cessation Support Via Mobile Phones: Current Knowledge and Future Directions. *Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco*, 19(3):379–383.

Navarro-Barrientos, J.-E., Rivera, D. E., and Collins, L. M. (2011). A dynamical model for describing behavioural interventions for weight loss and body composition change. *Mathematical and computer modelling of dynamical systems*, 17(2):183–203.

Ng, A. Y. and Russell, S. J. (2000). Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Nie, X., Brunskill, E., and Wager, S. (2021). Learning When-to-Treat Policies. *Journal of the American Statistical Association*, 116(533):392–409.

Pfammatter, A. F., Nahum-Shani, I., DeZelar, M., Scanlan, L., McFadden, H. G., Siddique, J., Hedeker, D., and Spring, B. (2019). SMART: Study protocol for a sequential multiple assignment randomized controlled trial to optimize weight loss management. *Contemporary Clinical Trials*, 82:36–45.

Pike-Burke, C. and Grünewälder, S. (2019). Recovering bandits. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 1265, pages 14122–14131. Curran Associates Inc., Red Hook, NY, USA.

Prince, M. J., Wu, F., Guo, Y., Gutierrez Robledo, L. M., O'Donnell, M., Sullivan, R., and Yusuf, S. (2015). The burden of disease in older people and implications for health policy and practice. *The Lancet*, 385(9967):549–562.

PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., Crawford, G. E., Jaffe, A. E., Pinto, D., Dracheva, S., Geschwind, D. H., Mill, J., Nairn, A. C., Abyzov, A., Pochareddy, S., Prabhakar, S., Weissman, S., Sullivan, P. F., State, M. W., Weng, Z., Peters, M. A., White, K. P., Gerstein, M. B., Amiri, A., Armoskus, C., Ashley-Koch, A. E., Bae, T., Beckel-Mitchener, A., Berman, B. P., Coetzee, G. A., Coppola, G., Francoeur, N., Fromer, M., Gao, R., Grennan, K., Herstein, J., Kavanagh, D. H., Ivanov, N. A., Jiang, Y., Kitchen, R. R., Kozlenkov, A., Kundakovic, M., Li, M., Li, Z., Liu, S., Mangravite, L. M., Mattei, E., Markenscoff-Papadimitriou, E., Navarro, F. C. P., North, N., Omberg, L., Panchision, D., Parikshak, N., Poschmann, J., Price, A. J., Purcaro, M., Reddy, T. E., Roussos, P., Schreiner, S., Scuderi, S., Sebra, R., Shibata, M., Shieh, A. W., Skarica, M., Sun, W., Swarup, V., Thomas, A., Tsuji, J., van Bakel, H., Wang, D., Wang, Y., Wang, K., Werling, D. M., Willsey, A. J., Witt, H., Won, H., Wong, C. C. Y., Wray, G. A., Wu, E. Y., Xu, X., Yao, L., Senthil, G., Lehner, T., Sklar, P., and Sestan, N. (2015). The PsychENCODE project. *Nature Neuroscience*, 18(12):1707–1712.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. (2017). Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, pages 147–163. PMLR.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Rivera, D. E., Pew, M. D., and Collins, L. M. (2007). Using Engineering Control Principles to Inform the Design of Adaptive Interventions: A Conceptual Introduction. *Drug and alcohol dependence*, 88(Suppl 2):S31–S40.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412.

Robins, J. M. (2000). Marginal Structural Models versus Structural nested Models as Tools for Causal inference. In Halloran, M. E. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, The IMA Volumes in Mathematics and its Applications, pages 95–133, New York, NY. Springer.

Robins, J. M. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. In Lin, D. Y. and Heagerty, P. J., editors, *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, Lecture Notes in Statistics, pages 189–326. Springer, New York, NY.

Rosenberg, E. S., Davidian, M., and Banks, H. T. (2007). Using mathematical modeling and control to develop structured treatment interruption strategies for HIV infection. *Drug and Alcohol Dependence*, 88 Suppl 2(Suppl 2):S41–51.

Rosenberger, W. F., Uschner, D., and Wang, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Statistics in Medicine*, 38(1):1–12.

Rosthøj, S., Fullwood, C., Henderson, R., and Stewart, S. (2006). Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Statistics in Medicine*, 25(24):4197–4215.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Ryan, J. C., Viana, J. N., Sellak, H., Gondalia, S., and O'Callaghan, N. (2021). Defining precision health: a scoping review protocol. *BMJ Open*, 11(2):e044663.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 29(4):640–661.

Shortreed, S. M., Laber, E., Scott Stroup, T., Pineau, J., and Murphy, S. A. (2014). A multiple imputation strategy for sequential multiple assignment randomized trials. *Statistics in Medicine*, 33(24):4202–4214.

Strecher, V. J., McClure, J. B., Alexander, G. L., Chakraborty, B., Nair, V. N., Konkel, J. M., Greene, S. M., Collins, L. M., Carlier, C. C., Wiese, C. J., Little, R. J., Pomerleau, C. S., and Pomerleau, O. F. (2008). Web-based smoking-cessation programs: results of a randomized trial. *American Journal of Preventive Medicine*, 34(5):373–381.

Sugiyama, M. (2015). *Statistical Reinforcement Learning: Modern Machine Learning Approaches*. Chapman and Hall/CRC.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. The MIT Press, Cambridge, Massachusetts, 2 edition.

Tao, Y. and Wang, L. (2017). Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics*, 73(1):145–155.

Tao, Y., Wang, L., and Almirall, D. (2018). Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *The Annals of Applied Statistics*, 12(3).

Tewari, A. and Murphy, S. A. (2017). From Ads to Interventions: Contextual Bandits in Mobile Health. In Rehg, J. M., Murphy, S. A., and Kumar, S., editors, *Mobile Health*, pages 495–517. Springer International Publishing, Cham.

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. E., and Tannir, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine*, 26(26):4687–4702.

Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology (Poznan, Poland)*, 19(1A):A68–77.

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2021). *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman & Hall/CRC, Boca Raton. OCLC: 1259526921.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477.

van der Laan, M. J. and Petersen, M. L. (2007). Statistical learning of origin-specific statically optimal individualized treatment rules. *The International Journal of Biostatistics*, 3(1):Article 6.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6:Article25.

Voils, C. I., Chang, Y., Crandell, J., Leeman, J., Sandelowski, M., and Maciejewski, M. L. (2012). Informing the dosing of interventions in randomized trials. *Contemporary Clinical Trials*, 33(6):1225–1230.

Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F. (2012). Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer. *Journal of the American Statistical Association*, 107(498):493–508.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. PhD, King's College, Cambridge University, Cambridge, UK.

Yu, C., Liu, J., Nemati, S., and Yin, G. (2023). Reinforcement Learning in Healthcare: A Survey. *ACM Computing Surveys*, 55(1):1–36.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics*, 68(4):1010–1018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhang, C., Chen, J., Fu, H., He, X., Zhao, Y.-Q., and Liu, Y. (2020). Multicategory Outcome Weighted Margin-based Learning for Estimating Individualized Treatment Rules. *Statistica Sinica*, 30:1857–1879.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26):3294–3315.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011). Reinforcement Learning Strategies for Clinical Trials in Non-small Cell Lung Cancer. *Biometrics*, 67(4):1422–1433.

Zhao, Y., Zhu, R., Chen, G., and Zheng, Y. (2020). Constructing dynamic treatment regimes with shared parameters for censored data. *Statistics in Medicine*, 39(9):1250–1263.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Journal of the American Statistical Association*, 110(510):583–598.

Zhou, M., Mintz, Y., Fukuoka, Y., Goldberg, K., Flowers, E., Kaminsky, P., Castillejo, A., and Aswani, A. (2018). Personalizing Mobile Fitness Apps using Reinforcement Learning. *CEUR Workshop Proceedings*, 2068.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017). Residual Weighted Learning for Estimating Individualized Treatment Rules. *Journal of the American Statistical Association*, 112(517):169–187.