International Statistical Review

International Statistical Review (2024) doi: 10.1111/insr.12583

Reinforcement Learning in Modern Biostatistics: Constructing Optimal Adaptive Interventions

Nina Deliu[1,2], Joseph Jay Williams[3] and Bibhas Chakraborty[4,5,6]

[1]*MEMOTEF Department, Sapienza University of Rome, Rome, Italy*
[2]*MRC – Biostatistics Unit, University of Cambridge, Cambridge, UK*
[3]*Department of Computer Science, University of Toronto, Toronto, Canada*
[4]*Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore*
[5]*Department of Statistics and Data Science, National University of Singapore (NUS), Singapore*
[6]*Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA*
Correspondence: Nina Deliu, MEMOTEF Department, Sapienza University of Rome, Rome, Italy.
Email: nina.deliu@uniroma1.it

Summary

In recent years, reinforcement learning (RL) has acquired a prominent position in health-related sequential decision-making problems, gaining traction as a valuable tool for delivering adaptive interventions (AIs). However, in part due to a poor synergy between the methodological and the applied communities, its real-life application is still limited and its potential is still to be realised. To address this gap, our work provides the first unified technical survey on RL methods, complemented with case studies, for constructing various types of AIs in healthcare. In particular, using the common methodological umbrella of RL, we bridge two seemingly different AI domains, dynamic treatment regimes and just-in-time adaptive interventions in mobile health, highlighting similarities and differences between them and discussing the implications of using RL. Open problems and considerations for future research directions are outlined. Finally, we leverage our experience in designing case studies in both areas to showcase the significant collaborative opportunities between statistical, RL and healthcare researchers in advancing AIs.

Key words: dynamic treatment regimes; just-in-time adaptive interventions; mobile health; multi-armed bandits; artificial intelligence; machine learning; reinforcement learning; optimal policy learning.

1 Introduction

In the era of big data and digital innovation, healthcare is going through a rapid and dramatic change process, transitioning from the *one-size-fits-all* standard to the tailored approach of *precision* or *personalised medicine* (Kosorok & Laber, 2019). Under this framework, the 'individual variability in genes, environment and lifestyle for each person' is taken into account in an effort to improve the ways we 'anticipate, prevent, diagnose and treat' a particular disease in a particular patient (Collins & Varmus, 2015). This paradigm encompasses a broad range of

scientific domains, ranging from genomics to advanced analytics and causal inference, all in support of a data-driven, yet patient-centric, approach for delivering personalised care.

One of the key methodological lines of research within the domain of personalised medicine is the development of *adaptive interventions* (AIs) (Almirall et al., 2014; Collins et al., 2004). The fundamental goal of AIs is to operationalise sequential decision-making by tailoring interventions to individuals, so as to offer guidance on how to adapt them to an individual's changing status and needs. In clinical practice, a typical situation is represented by a clinician who needs to use a set of treatment rules (i.e. a treatment regime) that recommend how to assign treatments or doses to patients based on their individual characteristics. These characteristics can include both baseline information (e.g. demographic data or pre-treatment clinical conditions) and evolving health status (e.g. responses to previous treatments). For example, for patients who do not improve on the first-line treatment over a prespecified period, the clinician may plan to increase the dose, according to a dose–response relationship, or change treatment in the case of a sensitive or drug-resistant patient. Due to changes in their health status, such a treatment regime is therefore *dynamic within* a person. To the patient, this sequence of treatments seems like standard treatment; to the clinician, it represents a series of prespecified decisions to make according to the patient's evolving history; and to the statistician, it constitutes an AI, alternatively known as *dynamic treatment regime* or *regimen* (DTR) (Chakraborty & Murphy, 2014; Lavori & Dawson, 2004; Murphy, 2003). The distinctive feature of AIs is their data-driven, adaptive approach guided by and oriented towards individual data. Clearly, an ambitious goal in AIs, or more specifically in DTRs, is how to construct the *optimal* DTRs, for example, treatment regimes that result in an optimal mean response or outcome. Such a question has a long history in statistics, and its study will occupy a central role in this work.

The traditional way of offering AIs to a patient mostly relies on rules created by experts, based on factors such as domain theory and empirical experience with similar patients. However, the recent advances and the widespread application of artificial intelligence and machine learning (ML) techniques (see, e.g. Deo, 2015; Oyebode et al., 2022; Rajkomar et al., 2019), have shed light on their ability to enable clinicians to quickly, efficiently and accurately identify the most appropriate course of action for their patients.

ML represents a hotspot in artificial intelligence, and health systems have recently tapped into its expanding potential to complement classical statistical tools and support clinical decision-making. There is no clear line between ML models and traditional statistical models (Beam & Kohane, 2018). Yet, it is widely acknowledged that sophisticated ML models such as deep learning models (Goodfellow et al., 2016) excel in learning, and automatically improving through experience, from the high-dimensional and heterogeneous data generated from modern clinical care. By matching a patient's characteristics to a computerised clinical knowledge base, such algorithms can suggest assessments or recommendations tailored to that patient's characteristics, even in very complex settings. Despite the potential to revolutionise decision-making, the success of ML in healthcare strongly depends on the efforts of the theoretical and methodological communities to unravel and elucidate their intrinsic mechanism and processes (often criticised for operating in a 'black box'). In turn, this may foster broader acceptance among clinicians and patients, thereby fostering greater confidence in the integration of ML technologies into clinical practice.

Among the existing ML paradigms (Bishop, 2006; Mohri et al., 2018), *reinforcement learning* (RL) (Bertsekas, 2019; Sugiyama, 2015; Sutton & Barto, 2018) offers a natural framework for the sequential decision-making problem encountered in AIs. In classical RL, a *learning agent* has to decide which of one or more *actions* to take when interacting with an *unknown environment*. Based on the feedback or *reward* received from the environment for the selected action(s), the agent learns how best to act to maximise the cumulative reward over time. This

is done by *trial-and-error*, that is, by observing and inferring from the environment after actions are taken. The RL framework is abstract and yet flexible enough to accommodate a variety of domains where the problem has a sequential nature (Chakraborty & Moodie, 2013; Gottesman et al., 2019); it does so by specifically characterising the environment's (or domain's) dynamics. In AIs, RL can be applied by regarding the alternative interventions as the actions to be chosen and the outcome of the intervention (e.g. patient's response) as the reward; patient's time-varying context and status represent the environment.

Within biostatistics, RL was first introduced as a data analysis tool to discover optimal DTRs in a variety of health domains including cancer (Goldberg & Kosorok, 2012; Zhao et al., 2009), weight loss management (Forman et al., 2019; Pfammatter et al., 2019), substance use (Chakraborty & Moodie, 2013; Murphy, Lynch, et al., 2007), mental health (Pike & Robinson, 2022) and so on. More recently, there seems to be an unprecedented interest in the application of RL in the rapidly expanding *mobile health* (mHealth) domain (Istepanian et al., 2006; Kumar et al., 2013; Kumar et al., 2017). The mHealth area refers to the use of mobile or wearable technologies to promote healthy behaviour changes in both clinical and non-clinical populations. A high-level goal in mHealth is to deliver efficacious *just-in-time adaptive interventions* (JITAIs) (Nahum-Shani et al., 2018) in response to the *in-the-moment* changes in an individual's internal state (e.g. health) and contextual state (e.g. location) (Kumar et al., 2017). The challenge in JITAIs is thus to provide 'the right individual with the right intervention', as well as 'the right intervention at the right time'. Notably, despite the relatively recent development of JITAIs compared with DTRs, research interest in both methodology and applications has substantially skewed towards JITAIs; we refer to Figure S1 in Supplementary Material A for a quantification of the volume of the literature. Given the increasing number of mHealth studies and in tandem the ongoing interest among statisticians in DTRs, integrating these two areas is a worthy objective. In the current article, we combine our methodological background with our experience in designing case studies in the above two areas to extensively review the state of the art of RL in AIs. To the best of our knowledge, this represents the first comprehensive survey of RL methods for developing DTRs as well as JITAIs in mHealth, informed by our experience with the challenges and successes of real-world applications. It complements and adds to the extensively surveyed DTR literature (see, e.g. Chakraborty & Moodie, 2013; Chakraborty & Murphy, 2014; Tsiatis et al., 2021), which we place together with JITAIs under the same AI umbrella.

We believe that there is ample scope for important practical advances in these areas, and with this survey, we aim to make it easier for theoretical and methodological researchers to join forces to assist healthcare discoveries by developing the next generation of methods for AIs in healthcare. We finally emphasise that we focus on healthcare and biostatistics due to the central role statisticians play there traditionally. Notwithstanding, the concepts we review for AIs extend far beyond: to education (Nahum-Shani & Almirall, 2019), policy making (Kasy & Sautmann, 2021) and other domains such as population research, where RL has recently been contextualised (Deliu, 2023).

The remainder of this work is structured as follows. In Section 2, we formally characterise the problem of AIs, providing a common framework for applications to DTRs and JITAIs, and explaining their similarities and differences. We then formalise the RL paradigm and its subclasses, relating it to the problems at hand and assimilating the different existing notations and terminologies into a coherent framework (Section 3). This provides a foundation for conducting research more easily in both methodological and applied aspects of AIs, enhancing communication and synergy between them. Section 4 offers a review of RL methods for developing AIs, expanding on DTRs with both finite- and indefinite-time horizons and JITAIs for mHealth. Insights on current methodological differences, along with their drivers, are discussed

in Section 4.3 and considerations for future research are provided in Section 4.4. Section 5 grounds Section 4 by illustrating the development and application of the presented methodology to two case studies. Section 6 concludes with some final remarks.

2 Adaptive Interventions in Healthcare

Adaptive interventions offer a vehicle to operationalise a sequential decision-making process over the course of a program or a condition, with the aim of optimising individual outcomes. Technically speaking, AIs are defined via explicit sequences of decision rules that pre-specify how the type, intensity and delivery of intervention options should be adjusted over time in response to individual progresses (Almirall et al., 2014; Nahum-Shani et al., 2018). The prespecified nature of AIs increases their replicability in research and enhances the assessment of their effectiveness (Nahum-Shani & Almirall, 2019).

The existing frameworks for formalising AIs (Almirall et al., 2014; Collins et al., 2004) are based primarily on four key components:

i The **decision points**, specifying the time points or time intervals at which a decision concerning intervention has been or has to be made; here, we assume a finite or countable number of times $t \in \mathbb{N} = \{0, 1, \ldots\}$;
ii The **decisions or intervention options** at each time t, that may correspond to different types, dosages (duration, frequency or amount; Voils et al., 2012), or delivery modes, as well as various tactical options (e.g. augment, switch and maintain); we denote them by $A_t \in \mathcal{A}_t$, where $\mathcal{A}_t$ is the decision or action space, generally discrete, at time t;
iii The **tailoring variable(s)** at each time t, say $X_t \in \mathcal{X}_t$, with $\mathcal{X}_t \subseteq \mathbb{R}^n$, $n \geq 1$, capturing individuals' baseline and time-varying information for personalising decision-making;
iv The **decision rules**, $\boldsymbol{d} = \{d_t\}_{t \geq 0}$, where, at each time t, d_t links the tailoring variable(s) X_t and potentially any other previous information deemed important to a specific decision or intervention $A_t \in \mathcal{A}_t$.

A common illustrative way to describe an AI is through schematics such as the one shown in Figure 1, where the 'if-then' statements clarify how the decision rule pre-specifies the intervention options under various conditions.

Because an AI adaptation is aimed at optimising individual outcomes, these play an essential role when defining an AI's components (Nahum-Shani & Almirall, 2019). In particular, we can distinguish between two types of individual outcomes:

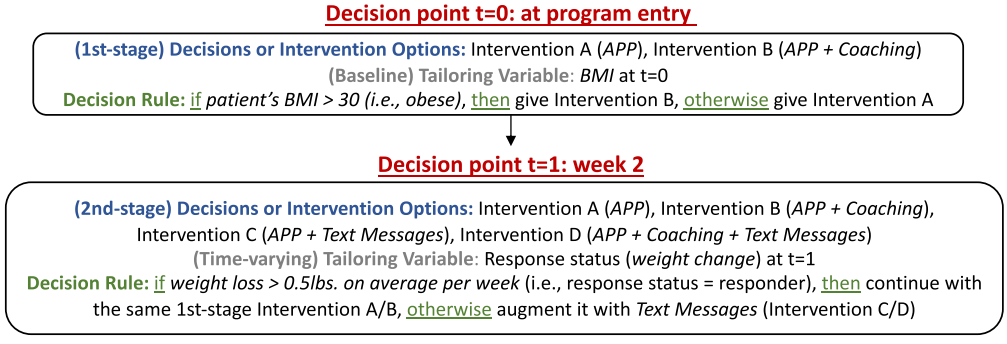

Figure 1. *Simplified schematic of a two-stage adaptive intervention and its key components, inspired by the weight loss management study in (Pfammatter et al., 2019).*

The **intermediate** or **proximal outcome(s)**, say $\{Y_t\}_{t>0}$, with $Y_t \in \mathcal{Y}_t$, that is, easily observable short-term outcome(s), expected to influence a longer-term outcome according to some mediation theory (MacKinnon et al., 2007);
The **final** or **distal outcome**, representing the long-term outcome of interest and the ultimate goal of the AI. To distinguish it from the intermediate stage-related outcomes $\{Y_t\}_{t>0}$, which may have a different meaning and nature, we denote it by $\overset{\infty}{Y}$.

Different AI problems would target different types of outcomes. For example, in some contexts, there may only be a distal (end-of-study) outcome $\overset{\infty}{Y}$ instead of multiple intermediate outcomes (see, e.g. Pelham et al., 2002): in that case, we will use the convention $Y_{T+1} \doteq \overset{\infty}{Y}$, with T being the study's last decision point or problem horizon. In other cases, only the intermediate outcomes will define the AI problem, while the final outcome will have no formal role in the decision-making problem. We also note that proximal outcomes can also be used as tailoring variables to guide later-stage decisions. In Figure 1, for example, the response status at time $t = 1$ represents both the proximal outcome targeted by the intervention at the decision point $t = 0$ and the tailoring variable at the decision point $t = 1$.

Development of AIs is based on the selection and integration of the aforementioned six components, taking into account their relationship. Ideally, this is informed and guided by domain theories, practical considerations, empirical evidence or some combinations thereof. Determining optimised decision rules typically involves more sophisticated data-driven statistical and ML tools, with RL recognised as a current state-of-the-art tool.

The term AI is interchangeably used with *adaptive treatment strategy* (Murphy, 2005a; Murphy, Collins, & Rush, 2007), *treatment policy* (Dawson & Lavori, 2012; Lunceford et al., 2002; Wahed & Tsiatis, 2006) and *dynamic treatment regime* or *regimen* (Chakraborty & Moodie, 2013; Laber, Lizotte, et al., 2014; Lavori & Dawson, 2004; Murphy, 2003), among others. However, given its more generic nature, we use the term AI to refer to a general framework for personalising interventions sequentially based on an individual's time-varying characteristics. This broader definition embraces a considerate number of applications, including non-healthcare (e.g. education; Nahum-Shani & Almirall, 2019) and the two healthcare domains of DTRs and JITAIs, which we cover below.

2.1 Dynamic Treatment Regimes

In medical research, DTRs define a sequence of treatment rules tailored to each individual patient based on their baseline and time-varying (dynamic) state. Traditionally, treatment assignment is based on *single-stage* decision-making. Specifically, one observes a set of baseline or pre-treatment information $X_0 \in \mathcal{X}_0$, based on which a treatment $A_0 \in \mathcal{A}_0$ is selected. The treatment rule, say d_0, is a mapping from $\mathcal{X}_0$ to $\mathcal{A}_0$. If more stages are involved, at each stage t, the treatment rule d_t is again an (independent) mapping from the stage-t information set $\mathcal{X}_t$ to a stage-t action space $\mathcal{A}_t$. Unlike single-stage protocols, where $\boldsymbol{d}_t = \{d_\tau : \mathcal{X}_\tau \to \mathcal{A}_\tau\}_{\tau=0,\ldots,t}$, DTRs explicitly incorporate the heterogeneity in treatment effect among individuals and *across time* within an individual, and regards $\boldsymbol{d}_t$ as a (nested) *multistage regime* with each $d_\tau, \tau = 0, \ldots, t$, depending on the individual evolving history of covariates, treatments and outcomes up to time τ; that is, $\boldsymbol{d}_t = \{d_\tau : \mathcal{H}_\tau \to \mathcal{A}_\tau\}_{\tau=0,\ldots,t}$, with $\mathcal{H}_\tau \doteq \mathcal{X}_0 \times \mathcal{A}_0 \times \mathcal{Y}_1 \times \ldots \times \mathcal{A}_{\tau-1} \times \mathcal{Y}_\tau \times \mathcal{X}_\tau$. As such, it provides an attractive framework for personalised treatments in longitudinal settings. Beyond personalisation, DTRs can identify and evaluate delayed effects, that is, effects that do not occur immediately after treatment but may affect a person or their disease later in time. It should also be noted that, by treating only those who show a need for treatment, DTRs hold the promise

of reducing non-compliance due to overtreatment or undertreatment (Lavori & Dawson, 2000). At the same time, they are attractive to public policy makers, allowing a better allocation of public and private funds (Murphy, 2003).

For developing DTRs, data sources include both longitudinal observational data, such as electronic health records (EHRs), and randomised studies, such as randomised-controlled trials (RCTs) and *sequential multiple assignment randomised trials* (SMARTs) (Lavori & Dawson, 2000; Murphy, 2005a). Although observational sources are much more common, SMARTs represent the current gold standard (Lei et al., 2012). A SMART is characterised by multiple stages of treatment, typically ranging from two to four, where each stage corresponds to one of the critical decision points. A concrete example is provided in Figure 2, which illustrates the first two stages of the weight loss management study in (Pfammatter et al., 2019). At study entry, all individuals are uniformly randomised to one of two first-line interventions: mobile app (APP) or APP + Coaching. Participants are assessed at weeks 2, 4, 8, and those 'responding' to their initial treatment (i.e. losing at least 0.5 lbs. on average per week) continue receiving the same treatment. As soon as an individual is classified as a 'non-responder', they are re-randomised to one of two augmentation tactics: modest augmentation (supportive text message; TXT) or vigorous augmentation (TXT + Coaching, or TXT + meal replacement (MR), depending on the first-stage treatment). Rerandomisation occurs only once per participant, with the newly assigned treatment continuing through the end of the study. Because different intervention options are considered for responders (continue) and non-responders (modest or vigorous augmentation), the response status is embedded as a tailoring variable. Such multistage restricted randomisation generates several DTRs *embedded* in the SMART; we refer to (Chakraborty & Moodie, 2013) for details on embedded regimes.

2.2 Just-In-Time Adaptive Interventions in Mobile Health

The ubiquitous use of mobile technologies has facilitated the development of a new area of health promotion in both clinical and non-clinical populations, known as mHealth (Istepanian et al., 2006). A key objective in mHealth is to deliver efficacious real-time AIs in response to rapid changes in individual circumstances, while avoiding overtreatment and its consequences on user engagement (e.g. low adherence to recommendations or discontinued usage of the mobile device). This specialised AI is termed just-in-time adaptive intervention (JITAI) (Nahum-Shani et al., 2018). JITAIs are nowadays gaining an increased popularity across various behavioural domains, spanning from physical activity (Figueroa et al., 2022; Hardeman et al., 2019) and weight management (Pfammatter et al., 2019) to addictive disorders (Garnett et al., 2019;

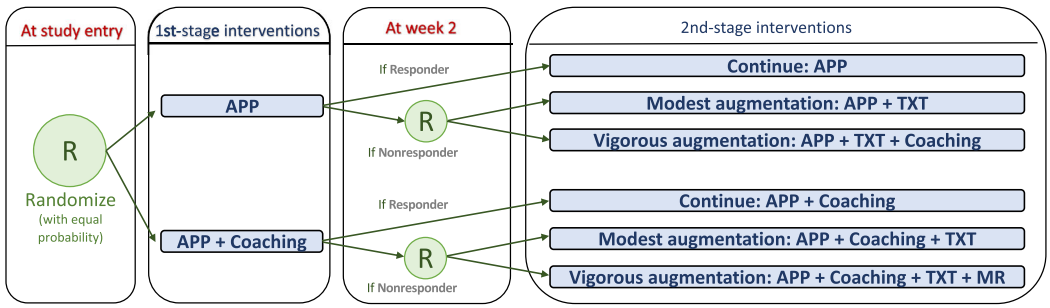

Figure 2. *Schematic of the weight loss SMART design in (Pfammatter et al., 2019).*

Goldstein et al., 2017; Naughton, 2017) and mental health (Kumar et al., 2024). Moreover, there has also been recent interest in leveraging JITAIs to enhance public health on a broader scale (Liu et al., 2023).

In mHealth, JITAIs refer to a sequence of decision rules that use continuously collected data through mobile technologies (e.g. wearable devices, accelerometers or smartphones) to adapt intervention components in real time in order to support behaviour change and to promote health. The peculiarity of JITAIs is that they deliver interventions according to the user's in-the-moment context or needs, for example, time, location or current activity, including considerations of whether and when the intervention is needed. Compared with DTRs, JITAIs are more flexible in terms of location and timing of interventions delivery. In fact, while the adaptation and delivery of a DTR usually take place at a pre-defined clinical appointment and under the direct guidance of a clinician, JITAIs often adapt and assign interventions as dictated by the mobile system or individual users, while they go about their daily lives in their natural environments. For this reason, unless otherwise designed, the time interval between decision points can vary significantly between and within subjects, dictated by randomness in individual needs and engagement with the mHealth device. Furthermore, 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 (as in the case of the *DIAMANTE* study, which will be shortly discussed and illustrated in Figure 3).

In JITAIs, the time between decision points is often too short to capture the (distal) clinical outcome of interest, and they rely on a weak surrogate, that is, the proximal outcome. Unlike DTRs—which target the distal outcome and may or may not have an intermediate (proximal) outcome—in JITAIs, proximal outcomes represent the direct and in-the-moment target of the intervention. The distal outcome is expected to improve only based on domain knowledge about its relationship with the proximal outcome, but is not formally included in the optimisation problem. We refer to Table 1 for a hand-to-hand comparison between JITAIs and DTRs under the AI framework.

Typical experimental designs for building JITAIs are represented by *factorial experiments* (Collins et al., 2009), or most notably, *micro-randomised trials* (MRTs) (Klasnja et al., 2015). In MRTs, individuals are randomised hundreds or thousands of times over the course of the study, and in a typical multicomponent intervention study, the multiple

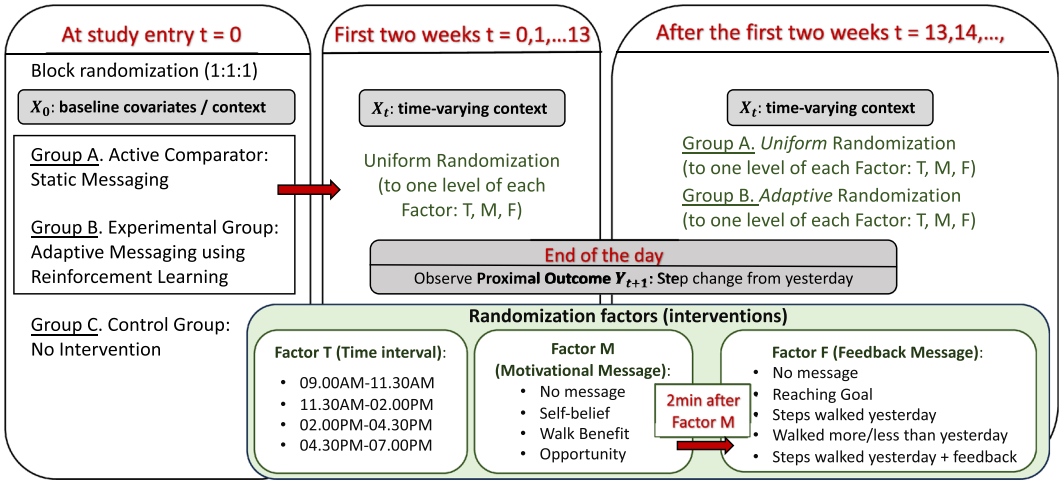

Figure 3. *Schematic of the DIAMANTE micro-randomised trial (Aguilera et al., 2020).*

Table 1. *Differences between DTRs and JITAIs in terms of the key characteristics defining an AI.*

	Type of adaptive intervention	
Characteristic	*DTRs*	*JITAIs in mHealth*
Data sources	RCTs, SMARTs, longitudinal observational data including EHRs, dynamical systems models	MRTs, RCTs, factorial designs, single-case experimental designs
AI component: (i) decision points $t \in \mathbb{N}$		
Number of decision points	In SMARTs, generally small (e.g. two to four); in EHRs (defined over indefinite horizons) an increased number is seen	Generally very large (hundreds or even thousands for each single unit)
Alignment of decision points across subjects	In SMARTs, these are expected to occur at some regular and fixed time points; in EHRs, these can reflect different protocols and a higher variability between individuals, with less regular patterns	Because decision points reflect user's specific needs and availability, in JITAIs, these are often random (as requested by the user)
Distance between decision points	Sufficiently long according to the expected time to capture a potential effect (including a delayed effect) of the intervention on the primary outcome of interest or a strong intermediate surrogate	Quite short according to the expected 'in-the-moment' effect of the intervention on the proximal outcome (e.g. every few minutes, hours or daily)
AI component: (ii) decisions or intervention options $A_t \in \mathcal{A}_t$, $t \in \mathbb{N}$		
Type of intervention	Mostly drugs or behavioural interventions	Generally behavioural interventions (e.g. motivational/feedback messages, coaching, reminders) with few exceptions (e.g. insulin adjustments)
Intervention delivery	Assigned by the care provider during an appointment or through digital devices	Assigned through digital/mobile devices according to an automatic algorithm or/and under care provider's guidance
AI component: (iii) tailoring variable(s) $X_t \in \mathcal{X}_t \subseteq \mathbb{R}^n$, $n \geq 1$, $t \in \mathbb{N}$		
Type of tailoring variable	Can include the full or partial history of baseline and time-varying patients' information. An external context can also be considered, but it has secondary relevance	Current users' information and any type of variable related to their momentary context (e.g. availability and weather), which plays a major role and can be very granular
AI component: (iv) decision rules $\boldsymbol{d} = \{d_t\}_{t \in \mathbb{N}}$		
Main strategy to optimise decision rules	• Offline methods for finite-horizon decision problems, with some exceptions (e.g. for EHRs-based DTRs an indefinite horizon may be considered) • While finite-horizon problems in general account for the full individual history over time, indefinite horizon problems assume a Markov structure	• Online methods over indefinite-time horizons • Considering the expected 'in-the-moment' effect of the intervention, typically, only the current or last observed information is accounted for, with a pre-dominant use of Markov, partially observed Markov, or simpler structures
AI component: (v)-(vi) outcomes		
Proximal outcome $Y_t \in \mathcal{Y}_t \in \mathbb{R}$, $t = 1, 2, \ldots$	• *Optional* short-term outcomes expected to impact the distal (long-term) outcome • While not being the primary target of the intervention, they may be part of the adaptation/optimisation process	• Short-term outcomes directly targeted by the intervention and expected to mediate the effect on the distal outcome • They guide the definition of just-in-time in the context of the identified problem, as well as the formulation of the adaptation strategy
Distal outcome $\overset{\infty}{Y}$	• The outcome directly and formally targeted by the intervention • The primary criterion that guides the adaptation/optimisation of the DTR, although intermediate outcomes are often part of the optimisation	• Long-term goal of a JITAI, expected to be influenced by an intervention through the mediating role of proximal outcomes (domain knowledge) • Typically, they do not guide the adaptation/optimisation of the learning strategy

components can be randomised concurrently, making micro-randomisation a form of a sequential full factorial design. The goal of these trials is to optimise mHealth interventions during the trial while offering a basis to assess the causal effects of each intervention component and to evaluate whether the intervention effects vary with time and/or with the individual contexts.

To better understand the characteristics and value of MRTs, let us now consider the *DIAMANTE* study for promoting physical activity, illustrated in Figure 3. In this study, the intervention components include whether or not to send a text message, which type of message to deliver, and at which time. The latter thus has a central role in this type of AI, as it defines the intervention set. The proximal outcome is the change in the number of steps a participant walked today from yesterday; and the context is given by a set of variables such as health information and study day. To assess the effectiveness of the optimised JITAI, users are assigned to different study groups: (A) a static (non-optimised) group, (B) an adaptive group based on RL and (C) a control group. In the two intervention groups, users are randomised every day to receive a combination of categories of the different intervention components, delivered within different time intervals. The adaptive RL-based optimised group will be briefly discussed in Section 5, after introducing the RL framework.

3 The Reinforcement Learning Framework

Generally speaking, RL is an area of ML concerned with determining optimal action selection policies in sequential decision-making problems (Bertsekas, 2019; Sutton & Barto, 2018). This framework is based on repeated interactions between a *decision maker* or *learning agent* and the *environment* it wants to learn about, to take better decisions or *actions*. Before characterising this process and formalising the RL problem(s), it is paramount to set out clearly the fundamental prerequisites that enable RL to solve decision-making problems such as in AIs with rigour.

3.1 A Preliminary Note: Causal Inference and Reinforcement Learning

While in this work our focus is primarily on RL, we note that this is neither a necessary nor generally a sufficient solution for building valid AIs. As we will mention in passing in Section 4.1.1, a variety of other traditional statistical approaches, mostly confined to the causal inference literature, exist and have a substantial relevance in the field. In fact, for developing AIs, one needs to assess the causal relationship between interventions and outcomes, thus, requiring an adequate framework for causality.

Causal inference provides a set of tools and principles that allows one to combine data and causal assumptions about the environment to reason with questions of counterfactual nature. Through considerations on study designs, estimation strategies and certain fundamental assumptions (e.g. no unmeasured confounders), it provides the building blocks that enable researchers to draw causal conclusions based on the observed data. On a different tangent, RL is concerned with efficiently finding a policy that optimises an objective function (e.g. the expected cumulative reward) in interactive and uncertain environments. In practice, despite being causal by nature—any system looking to advise on interventions in some way quantifies their effects—the classical RL does not conduct causal inference. We can think of at least two reasons. First, RL practitioners often consider problems in which the data are unconfounded (e.g. robotics), because these are collected through direct interactions with a relatively well-understood environment, governed by physical laws, and actions are taken by the learning agent depending only on the data available (experimental data). To illustrate, this is the case of current JITAI practices. Second, and most importantly, the fundamental problem of RL, rather

than dealing with causal effects estimation, is oriented towards causal-decision making. We note that the two are not the same, and counterintuitively, accurate estimation is not essential for accurate decision-making (Fernández-Loría & Provost, 2022). While these two areas have evolved independently over different aspects of the same building block and with no interaction between them, disciplines such as AIs can be developed only under an integrated framework that permits causal conclusions.

Our attention in the current work is devoted to RL rather than causal inference, and we point the readers to the seminal works of Neyman and Rubin (Neyman, 1923; Rubin, 1974) for the potential outcomes framework, and to Pearl (Pearl, 2009) for the causal graphical model perspective. For a comprehensive treatment of both, we refer to (Hernan & Robins, 2023). Furthermore, recent attempts in the ML community have worked towards a unified framework called causal RL, which embeds the causal graphical approach within sample efficient RL algorithms (Zhang & Bareinboim, 2020). For simplicity of exposition, in this work, we assume that the main assumptions of causal inference (see, e.g. Chakraborty & Murphy, 2014) hold and that the conditional distributions of the observed data are the same as the conditional distributions of the potential outcomes, given the assigned treatment. It follows that RL can operate in a simplified causal inference problem (in which actions are unconfounded) and that optimal AIs may be obtained using the observed data.

3.2 Formalisation of the General Reinforcement Learning Problem

Consider a discrete time space indexed by $t \in \mathbb{N} = \{0, 1, \ldots,\}$. In RL, at each decision time point or simply time t, an agent faces a decision-making problem in an unknown environment. After receiving some representation of the environment's *state* or *context*, say $X_t \in \mathcal{X}_t$, it selects an *action*, denoted by A_t, from a set of admissible actions $\mathcal{A}_t$. As a result, one step later, the environment responds to the agent's action by making a transition into a new state $X_{t+1} \in \mathcal{X}_{t+1}$ and (typically) providing a numerical *reward* $Y_{t+1} \in \mathcal{Y}_{t+1} \subset \mathbb{R}$. By repeating this process over time, the result is a trajectory of states visited, actions pursued and rewards received. In a medical context, this trajectory can be viewed as the individual *history* (of covariates, treatments and responses to treatments) of a patient over time. Note that in some settings there may be only one terminal reward (or a final outcome, e.g. overall survival or school performance at the end of the study; Pelham et al., 2002); in this case, rewards at all previous time points are taken to be 0. In other settings (e.g. multi-armed bandits; Section 3.3.2), states may be ignored, thus leading to a trajectory of actions and rewards only.

Define $\mathbf{X_t} \doteq (X_0, \ldots, X_t)$, $\mathbf{A_t} \doteq (A_0, \ldots, A_t)$, $\mathbf{Y_{t+1}} \doteq (Y_1, \ldots, Y_{t+1})$, 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 realisations, respectively. Define *history* $\mathbf{H_t}$ as all the information available at time t prior to decision A_t, that is, $\mathbf{H_t} \doteq (\mathbf{A_{t-1}}, \mathbf{X_t}, \mathbf{Y_t})$; similarly $\mathbf{h_t}$. The history $\mathbf{H_t}$ at time t belongs to the product set $\mathcal{H}_t = \mathcal{X}_0 \times \prod_{\tau=1}^{t} \mathcal{X}_\tau \times \mathcal{A}_{\tau-1} \times \mathcal{Y}_\tau$. Note that, by definition, $\mathbf{H_0} = X_0$. We assume that each longitudinal history is sampled independently according to a distribution $P_{\boldsymbol{\pi}}^{\text{Full}-\text{RL}}$ (with the superscript clarified later in Section 3.3), given by

$$P_{\boldsymbol{\pi}}^{\text{Full}-\text{RL}} \doteq p_0(x_0) \prod_{t \geq 0} \pi_t(a_t|\mathbf{h_t}) p_{t+1}\left(x_{t+1}, y_{t+1}|\mathbf{h_t}, a_t\right), \tag{1}$$

where

- p_0 is the probability distribution of the initial state X_0.
- $\boldsymbol{\pi} \doteq \{\pi_t\}_{t \geq 0}$ represents the *exploration policy* that 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 $\mathcal{A}_t$, that is, $\pi_t(\cdot|\mathbf{h_t})$. The conditioning symbol '|' in $\pi_t(\cdot|\mathbf{h_t})$ reminds us that the exploration policy defines a probability distribution over $\mathcal{A}_t$ for each $\mathbf{h_t} \in \mathcal{H}_t$. 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* and they completely characterise 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 $\mathcal{X}_t \times \mathcal{Y}_t$, that is, $p_t(\cdot, \cdot|\mathbf{h_{t-1}}, a_{t-1})$.

  At each time t, the transition probability distribution $p_{t+1}(x_{t+1}, y_{t+1}|\mathbf{h_t}, a_t)$ gives rise to

i $p_{t+1}(x_{t+1}|\mathbf{h_t}, a_t)$, the *state-transition probability distribution*, representing the probability of moving to state x_{t+1} having observed history $\mathbf{h_t}$ and taking action a_t;

ii $p_{t+1}(y_{t+1}|\mathbf{h_t}, a_t, x_{t+1})$, the *immediate reward distribution*, specifying the reward Y_{t+1} after transitioning to x_{t+1} with action a_t.

Generally, in DTRs, the immediate reward Y_{t+1} is conceptualised as a known function of the history $\mathbf{H_t}$, the current selected action A_t and the new state X_{t+1}; that is, conditional on $\mathbf{H_t}$, the reward function is deterministic and Y_{t+1} is uniquely determined. To give a concrete example, one can think of a dose-finding trial, where the level of toxicity is one of the state variables, among others. In this setting, at each time t, the immediate reward Y_{t+1} of a patient with history $\mathbf{H_t}$ and administered dose A_t could be defined as a binary variable assuming value -1 if the observed toxicity level (X_{t+1}) is higher than a certain prespecified threshold, and 0 otherwise.

The cumulative sum (often time-discounted) of immediate rewards is termed *return*, say $\mathbf{R_t}$, and is given by

$$\mathbf{R_t} \doteq Y_{t+1} + \gamma Y_{t+2} + \gamma^2 Y_{t+3} + \ldots = \sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1}, \tag{2}$$

for $t \in \mathbb{N}$. 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 $\gamma < 1$, the potential infinite sum in Equation (2) has a finite value as long as the reward sequence $\{Y_{\tau+1}\}_{\tau \geq t}$ is bounded. If $\gamma = 0$, the agent is *myopic* in being concerned only with maximising the immediate reward, that is, $\mathbf{R_t} = Y_{t+1}$; this is often the case of the multi-armed bandit framework (see Section 3.3.2). If $\gamma = 1$, the return is *undiscounted* and it is well defined (finite) as long as the time horizon is finite, that is, $t \in [0, T]$, with $T < \infty$ (Sutton & Barto, 2018). If T is fixed and known in advance, for example, in clinical trials, the agent faces a *finite-horizon* problem; if T is not prespecified and can be arbitrarily large (the typical case of EHRs), but finite, we call it an *indefinite-horizon* problem; finally, we use the term *infinite-horizon* problem when $T = \infty$. In this case, we need $\gamma \in (0, 1)$ to ensure a well-defined return. As preliminarily outlined in Table 1, DTRs mainly deal with finite-horizon problems (exception made for EHRs), while JITAIs involve indefinite-horizon problems.

3.2.1 *Online and offline reinforcement learning*

Solving an RL task means learning an optimal way to choose the set of actions, or learning an *optimal policy*, so as to maximise the expected future return. This process may follow two learning strategies: *online* or *offline*. In online learning, an agent learns and improves/optimises the exploration policy $\boldsymbol{\pi}$ while following it, that is, from experiences sampled directly from $\boldsymbol{\pi}$. The policy $\boldsymbol{\pi}$ represents both the data-generating policy and the *target policy*, say $\boldsymbol{d}$, the agent wants to learn about ($\boldsymbol{\pi} = \boldsymbol{d}$). However, in many decision problems, for example, in

observational settings, the agent has to learn from previously collected data. In this case, the target policy $\boldsymbol{d}$ is learned from samples collected with a policy that can be either the exploration policy (when known, e.g. in randomised studies), or, more generally, an observed or *behaviour policy* (that is, $\boldsymbol{\pi} \neq \boldsymbol{d}$). In the AI space, the DTR literature has predominantly focused on offline learning strategies (typically from observational data), while the mHealth domain has often adopted online RL (under a randomised setting).

In a general RL problem, where the exploration/behavioural policy and the target policy of interest $\boldsymbol{d}$ can differ, the goal is to find an optimal policy $\boldsymbol{d}_t^* \doteq \{d_t^*\}_{\tau \geq t}$ at any time t, such that

$$\boldsymbol{d}_t^* = \arg\max_{\boldsymbol{d}_t} \mathbb{E}_{\boldsymbol{d}}[\mathbf{R_t}] = \arg\max_{\boldsymbol{d}_t} \mathbb{E}_{\boldsymbol{d}}\left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1}\right], \tag{3}$$

where the expectation is meant with respect to a trajectory distribution analogous to Equation (1), say $P_{\boldsymbol{d}}$, with $\boldsymbol{\pi}$ replaced by a general target policy $\boldsymbol{d}$.

To estimate optimal policies, various methods have been developed so far in the RL literature (see Sugiyama, 2015; Sutton & Barto, 2018, for an overview). A traditional approach is through *value functions*, which are classified into two main types: (i) *state-value* or simply *value* functions, representing how good it is for an agent to be in a given state, and (ii) *action-value* functions, indicating how good it is for the agent to perform a given action in a given state.

More specifically, the time-t *state-value function* of policy $\boldsymbol{d}$ gives us the expected return of following policy $\boldsymbol{d}$ from time t onward, conditional on history $\mathbf{h_t}$. Formally, we denote it by $V_t^{\boldsymbol{d}}: \mathcal{H}_t \to \mathbb{R}$ and define it as

$$V_t^{\boldsymbol{d}}(\mathbf{h_t}) \doteq \mathbb{E}_{\boldsymbol{d}}[\mathbf{R_t} | \mathbf{H_t} = \mathbf{h_t}] = \mathbb{E}_{\boldsymbol{d}}\left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1} \middle| \mathbf{H_t} = \mathbf{h_t}\right], \quad \forall \mathbf{h_t} \in \mathcal{H}_t, \quad \forall t \in \mathbb{N}. \tag{4}$$

To ensure that the conditional expectation in $V_t^{\boldsymbol{d}}(\mathbf{h_t})$ is well defined, each history $\mathbf{h_t} \in \mathcal{H}_t$ should have a positive probability to occur, that is, $\mathbb{P}(\mathbf{H_t} = \mathbf{h_t}) > 0$. Note that, by definition, at time $t = 0$, $V_0^{\boldsymbol{d}}(\mathbf{h_0}) \doteq V_0^{\boldsymbol{d}}(x_0)$; while for the terminal time point, if any, the state-value function is 0.

It is interesting to note that value functions define a partial ordering over policies with insightful information on the optimal ones. In fact, according to the definition of optimal policies given in Equation (3), a policy $\boldsymbol{d}$ is better than or equal to a policy $\boldsymbol{d}'$ if its expected return is greater than or equal (denoted as $\succeq$) to that of $\boldsymbol{d}'$ for all possible histories. Equivalently, $\boldsymbol{d} \succeq \boldsymbol{d}'$ if and only if $V_t^{\boldsymbol{d}}(\mathbf{h_t}) \geq V_t^{\boldsymbol{d}'}(\mathbf{h_t})$ for all $\mathbf{h_t} \in \mathcal{H}_t$. As a result, optimal policies share the same (optimal) value function. Efficient estimation of the value function represents one of the most important components of almost all RL algorithms, with a central place in the decision-making paradigm. In DTRs, for example, evaluating the value function of a treatment regime is equivalent to evaluating the average outcome if the estimated treatment rule were to be applied to a population with the same characteristics (state or history) in the future (Zhu et al., 2019). Comparing the estimated value functions of different candidate treatment regimes offers a way to understand which regime may offer the greatest expected outcome.

Similar insights are given by the *action-value function*. The time-t *action-value function* for policy $\boldsymbol{d}$, denoted by $Q_t^{\boldsymbol{d}}$, where 'Q' stands for 'Quality', is the expected return when starting from history $\mathbf{h_t}$ at time t, taking an action a_t and following the policy $\boldsymbol{d}$ thereafter. Formally, $\forall t \in \mathbb{N}$, $Q_t^{\boldsymbol{d}}: \mathcal{H}_t \times \mathcal{A}_t \to \mathbb{R}$ is defined as

$$Q_t^{\boldsymbol{d}}(\mathbf{h_t}, a_t) \doteq \mathbb{E}_{\boldsymbol{d}}[\mathbf{R_t} \,|\, \mathbf{H_t} = \mathbf{h_t}, A_t = a_t] = \mathbb{E}_{\boldsymbol{d}}\left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1} \middle| \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right], \quad \forall a_t \in \mathcal{A}_t, \quad \forall \mathbf{h_t} \in \mathcal{H}_t. \tag{5}$$

This is also known as *Q-function*, and as in Equation (4), $\mathbf{H_t}$ and A_t are such that $\mathbb{P}(\mathbf{H_t} = \mathbf{h_t}) > 0$ and $\mathbb{P}(A_t = a_t) > 0$.

At time t, the *optimal value function* $V_t^* \doteq V_t^{\boldsymbol{d}^*}$ yields the largest expected return for each history with any policy $\boldsymbol{d}$, and the *optimal Q-function* $Q_t^* \doteq Q_t^{\boldsymbol{d}^*}$ yields the largest expected return for each history-action pair with any policy $\boldsymbol{d}$, that is,

$$Q_t^*(\mathbf{h_t}, a_t) \doteq \max_{\boldsymbol{d_t}} Q_t^{\boldsymbol{d}}(\mathbf{h_t}, a_t), \quad \forall \mathbf{h_t} \in \mathcal{H}_t, \forall a_t \in \mathcal{A}_t; \tag{6}$$

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

Because an optimal action-value function is optimal for any fixed $\mathbf{h_t} \in \mathcal{H}_t$, it follows that the optimal policy at time t must also satisfy

$$d_t^*(\mathbf{h_t}) \in \arg\max_{a_t \in \mathcal{A}_t} Q_t^*(\mathbf{h_t}, a_t). \tag{8}$$

A fundamental property of the value functions used throughout RL is that they satisfy particular recursive relationships, known as *Bellman* equations (Bellman, 1957). For any policy $\boldsymbol{d}$, the following consistency condition, expressing the relationship between the value of a state and the values of the successor states, holds:

$$V_t^{\boldsymbol{d}}(\mathbf{h_t}) = \mathbb{E}_{\boldsymbol{d}}\left[Y_{t+1} + \gamma V_{t+1}^{\boldsymbol{d}}(\mathbf{h_{t+1}}) \,|\, \mathbf{H_t} = \mathbf{h_t}\right], \quad \forall \mathbf{h_t} \in \mathcal{H}_t, \quad \forall t \in \mathbb{N}. \tag{9}$$

Based on this property and Equations (6)–(7), at each time t, $\forall \mathbf{h_t} \in \mathcal{H}_t$ and $\forall a_t \in \mathcal{A}_t$, with discrete state and action spaces, the following rules, known as *Bellman optimality* equations (9), are satisfied:

$$V_t^*(\mathbf{h_t}) = \mathbb{E}\left[Y_{t+1} + \gamma V_{t+1}^*(\mathbf{h_{t+1}}) | \mathbf{H_t} = \mathbf{h_t}\right]; \tag{10}$$

$$Q_t^*(\mathbf{h_t}, a_t) = \mathbb{E}\left[Y_{t+1} + \gamma \max_{a_{t+1} \in \mathcal{A}_{t+1}} Q_{t+1}^*(\mathbf{h_{t+1}}, a_{t+1}) | \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right]. \tag{11}$$

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. This property allows for the 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 for solving the Bellman equation, also referred to as the DP equation or *optimality equation* (Sutton & Barto, 2018). In infinite- and indefinite-horizon problems, using 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 focus on time-homogeneous Markov processes to eliminate the dependence of value functions on t (see, e.g. Ertefaie & Strawderman, 2018; Luckett et al., 2020), or revisit the Bellman optimality equation (Zhou et al., 2022).

3.3 Formalisation of Specific Reinforcement Learning Problems

The RL problem can be posed in a variety of different ways depending on the assumptions about the level of knowledge initially available to the agent. The framework is abstract yet flexible enough to be applied to many different (sequential) problems by specifically characterising

the state and action spaces, the reward function, and other general domain (or environment) aspects, such as the time horizon or the dynamics of the process. The general framework introduced in Section 3.2 does not make any simplifying assumptions about the dependency between rewards, actions and states: by carrying over all the available history from $t = 0$, it considers a full dependency between them. We name this framework *full reinforcement learning* (full-RL).

Often, specific domains of application may have an underlying theory about the potential relationships between the key elements of an RL problem. For example, one may find it plausible to ignore the overall history and consider only the current state in the decision-making process. Furthermore, in some applied problems (e.g. indefinite-horizon problems), a full-RL formalisation may be infeasible and/or intractable for both optimisation and inference purposes. Thus, some forms of simplification in the distribution of the longitudinal histories may be needed. For example, in JITAIs, the 'just-in-time' nature of decision-making requires a computationally feasible estimation and application of the decision rule continuously in time.

Common examples of specific formalisations of an RL problem include *Markov decision processes* (MDPs) and *multi-armed bandit* (MAB) or contextual MAB problems. Although we discuss the MAB problem as a subclass of—or a special way of formalising—the RL problem (as in Sutton & Barto, 2018), we want to point out that some key researchers in the domain (see, e.g. Lattimore & Szepesvári, 2020) distinguish between the two. According to them, RL is mostly associated with ML, whereas MABs are with mathematics. One driver of this choice may be related to the major focus and attention to theoretical guarantees on *regret* bounds that MAB algorithms seek to satisfy.

In what follows, we illustrate these two specific formalisations, starting with the MDPs, the main framework for indefinite-horizon DTR problems. A graphical illustration of the different settings is given in Figure 4.

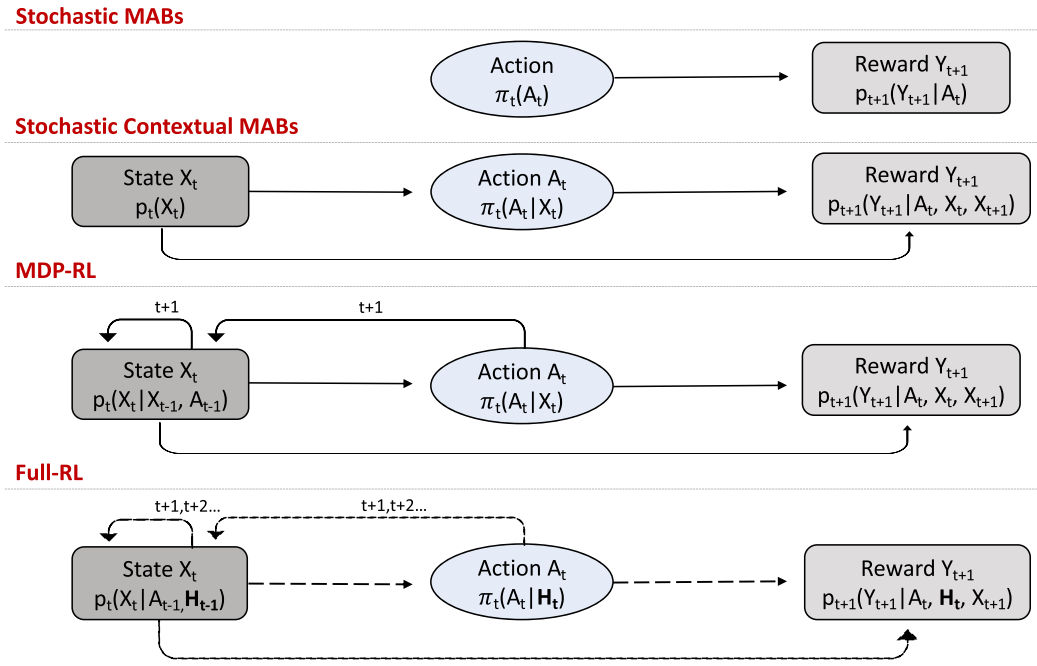

Figure 4. *Graphical representation of the states, actions and rewards relationship in a full-RL, MDP-based RL (MDP-RL), and stochastic (both contextual and context free) MAB. Solid and dashed lines indicate a direct and indirect (e.g. time-delayed) effect, respectively.*

3.3.1 *Markov decision processes*

An MDP is a stochastic process used to define the dynamics of an environment and to model the interaction between the agent and the environment. It provides a convenient mathematical framework for modelling decision-making in situations where the environment is deemed to evolve according to the *Markov model* (Puterman, 1994). Notably, it is the most common setting assumed in RL (Van Otterlo & Wiering, 2012).

What distinguishes an MDP-based RL (MDP-RL) from the full-RL framework is the environment's random memoryless characteristic. More specifically, assuming that the current state X_t contains all the information of the past history $\mathbf{H_{t-1}}$ relevant to future predictions, it allows us to ignore the past when modelling future states and rewards. This property, known as *Markov property*, leads to a low-dimensional representation of the past, exemplifying the trajectory distribution in Equation (1) as follows:

$$\begin{aligned} P_{\boldsymbol{\pi}}^{\mathrm{MDP}} &\doteq p_0(x_0) \prod_{t \geq 0} \pi_t(a_t|x_t) p_{t+1}\left(x_{t+1}, y_{t+1}|x_t, a_t\right) \\ &= p_0(x_0) \prod_{t \geq 0} \pi_t(a_t|x_t) p_{t+1}(x_{t+1}|x_t, a_t) p_{t+1}\left(y_{t+1}|x_t, x_{t+1}, a_t\right). \end{aligned}$$

Note that under the Markov property, the agent's decisions can be entirely determined based on the current information only, as it fully determines the environment's transition-probability distributions, that is, $p_{t+1}(\cdot\,, \cdot\,|\mathbf{H_t}, A_t) = p_{t+1}(\cdot\,, \cdot\,|X_t, A_t)$, for all t. When the transition probabilities $\{p_{t+1}\}_{t \geq 0}$ are also time independent, that is, $p_{t+1} = p$, for all t, the process is called *time-homogeneous* or *stationary* MDP. In light of this additional assumption, states, rewards and actions are now time independent, given the information of previous time points. In the context of DTRs as well as JITAIs, time-homogeneous MDPs were proposed in indefinite-time horizons, as they simplify the problem by working with time-independent quantities, which do not require a backward induction strategy (see Section 4.1.3).

While both full-RL and MDP-RL are typically formulated as problems with states, actions, rewards and transition rules that depend on previous states, an exception is made for MABs, whose original formulation can be viewed as a *stateless* variant of RL. In a typical MAB problem, either the actions and the rewards are not associated with states or they are assumed to depend only on the current state. This feature enables faster learning in settings such as JITAIs where RL is continuously implemented in an *online* fashion. This aspect will be discussed in more detail in Section 4.3.

3.3.2 *Multi-armed bandits*

MAB problems, often identified as a special subclass of RL (Sutton & Barto, 2018), have a long history in statistics. They were introduced in 1933 by (Thompson, 1933) and extensively studied under the heading *sequential design of experiments* (Lai & Robbins, 1985; Robbins, 1952).

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 maximise the total expected reward over time. Each of the K choices (i.e. *arms* or actions) provide 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 optimising its decisions based on acquired knowledge up to time

t (*exploitation*) and acquiring new knowledge about the expected rewards of the other actions (*exploration*).

MAB strategies were originally proposed to solve stateless problems, in which the reward depends uniquely on actions. Subsequently, a 'stateful' variant of MABs, named *contextual* MAB (C-MAB), in which actions are associated with some state, or *context*, was introduced. However, unlike full-RL and MDP-RL, in contextual MABs, actions do not have any effect on the next states. In addition, generally, there are no transition rules from one state to another in subsequent times. This implies that states, actions and rewards can be treated as a set of separate events over time. The most typical assumption is that contexts $\{X_t\}_{t \in \mathbb{N}}$ are independent and identically distributed (IID) 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 IID property holds as well. Under this assumption, one can be completely myopic and ignore the effect of an action on the distant future in searching for a good policy. This problem is better known as *stochastic* MABs, in contrast to *adversarial* MABs (Lattimore & Szepesvári, 2020), in which no independence assumptions are made on the sequence of rewards. In stochastic contextual MABs, and further in the context-free MAB problem, the trajectory distributions are simplified as follows:

$$\begin{aligned} P_{\boldsymbol{\pi}}^{\mathrm{C-MAB}} &\doteq p_0(x_0) \prod_{t \geq 0} \pi_t(a_t|x_t) p_{t+1}\left(x_{t+1}, y_{t+1}|x_t, a_t\right) = p_0(x_0) \prod_{t \geq 0} \pi_t(a_t|x_t) p_{t+1}(x_{t+1}) p_{t+1}\left(y_{t+1}|x_t, x_{t+1}, a_t\right); \\ P_{\boldsymbol{\pi}}^{\mathrm{MAB}} &\doteq \prod_{t \geq 0} \pi_t(a_t) p_{t+1}\left(y_{t+1}|a_t\right). \end{aligned} \tag{12}$$

Note that, because the effect of an action in the stochastic MAB is in-the-moment, the bandit problem is formally equivalent to a one-step/state MDP, wherein the states progression is not taken into account. Thus, compared with MDP-RL and full-RL, MABs provide a simplified structure of the relationships between the components of RL within time. For a graphical summary, see Figure 4.

As in the general RL problem, the goal of an MAB problem is to select the optimal arm at each time t so as to maximise the expected return, alternatively (and with a slightly different nuance) expressed in the bandit literature in terms of minimising the *total regret*. Indeed, in (online) real-world problems, until we can identify the best (unique) arm, we need to make repeated trials by pulling the different arms. The loss that we incur during this learning phase (i.e. the time spent for learning the best arm) represents what is called regret, that is, how much we regret not picking the best arm. Formally, denoted by $A_t^* \doteq \mathrm{argmax}_{a_t \in \mathcal{A}} \mathbb{E}(Y_{t+1}|, |X_t = x_t|, , A_t = a_t)$ 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, that is,

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

Given a horizon T, the goal of the learner is to minimise the total regret given by $\mathrm{Reg}(T) \doteq \sum_{t=0}^{T} \Delta(A_t)$. Note that the agent may not know ahead of time how many time points T are to be played. Therefore, the goal is to perform well not only at the final time point T, but also during the learning phase. For example, in a dose-finding problem as the one mentioned in Section 3.3.1, the objective may not only be to minimise the sum of toxicities over time, but also to ensure that these toxicities have a proper upper limit—thus, limiting extremely harmful adverse events—uniformly over time. For this reason, as we will see later in Section 4.2, theoretical works on regret bounds occupy a central place in the bandit literature.

3.4 Reinforcement Learning and Adaptive Interventions: A Joint Overview

So far, we have introduced the RL as a mathematical framework for sequential decision-making problems and discussed its characterisation in illustrative AI examples of interest. Before diving deep into the rich literature of existing RL methods for building (optimised) AIs, we provide the reader with a joint overview of the different problems, which notably share the same key elements and a common optimisation objective. As such, they can be unified under a unique formal framework and solved with techniques developed under the RL paradigm.

Table 2 outlines the terminologies of reference in each setting, with a unified notation adopted from the general RL. Note that, while we report only the most common terminology employed in each setting, lexical borrowing is widely used across the different theoretical and applied domains. To illustrate, the term 'treatment policy', or just 'policy' is often used in place of 'treatment regime' in the DTR literature. Also note that, in general, the terminology adopted in a specific application is guided by the RL method and framework used in that application; see, for example the similarity between the terms used in JITAIs and MABs such as 'contextual variables' and 'context' (i.e. the state of the environment). Both contextual and tailoring variables represent the set of baseline and time-varying information that is used to personalise decision-making. Alternative terms such as covariates or features (which we use with slightly different meaning, as we discuss in Section 4.2.1) are also common. To help the reader navigate the different terminologies, an extended version of Table 2, detailing all the notation and acronyms used in the main manuscript, is provided in Supplementary Material B.

We anticipate that most (if not all) of the methods to construct JITAIs would generally belong to the MAB class, although the applied literature commonly refers to it with the generic 'reinforcement learning' name (see, e.g. Figueroa et al., 2021; Liao et al., 2020; Yom-Tov et al., 2017). In DTRs, the predominant class of methods is full-RL, followed by MDP-RL proposed specifically for indefinite-horizon (e.g. EHR-based) DTR problems. In fact, the underlying theory of DTRs—characterised by potential delayed or carried-over effects of treatment over time—and the importance of the evolving history of a patient for predicting future outcomes requires accurate consideration of information from previous time points. Generally, the meaningful relationship between the different variables of a patient's history does not allow simplifying or ignoring the (state-)transition rules, making full-RL (and occasionally MDP-RL) the ideal option. On the other hand, the behavioural theory of a momentary effect of an intervention on the proximal outcome makes MABs a suitable framework in mHealth settings. In addition, the reduced computational burden from carrying through all the historical information allows MAB strategies to be applied continuously in time, for example, every hour, and efficiently construct JITAIs.

Table 2. *Notation and terminology of reference of the key elements in RL, MAB, DTR and JITAI problems.*

Notation	*Terminology*			
	RL	*MABs*	*DTRs*	*JITAIs*
i	Trajectory	Trajectory	Patient	User
t	Time point	Round, time point	Stage, interval, time point	Time point
X	State	Context	Tailoring variable	Contextual variable
A	Action	Arm	Treatment, intervention	Intervention option
Y	Reward	Reward	Intermediate outcome	Proximal outcome
$\mathbf{H}$	History	History, filtration	History	History
$\boldsymbol{\pi}$, $\boldsymbol{d}$	Policy	Policy	(Dynamic) treatment regime	Policy
$\boldsymbol{\pi}^*$, $\boldsymbol{d}^*$	Optimal policy	Optimal policy	Optimal DTR	Optimal policy

4 A Survey of Reinforcement Learning Methods for Adaptive Interventions

Methodology for constructing *optimal* AIs, that is, the ones that, if followed, would yield the most favourable (typically long-term) mean outcome, is of considerable interest within the domain of precision medicine, and comprises a large body of research within theoretical and applied sciences (Chakraborty & Moodie, 2013; Kosorok & Laber, 2019; Laber, Lizotte, et al., 2014). Although their relevance has been long documented within statistics and causal inference (see Section 4.1.1), recently it has generated a lot of interest within the computer science and engineering communities, due to the similarity between the mathematical formalisation of AIs and the RL framework.

4.1 Methods for Dynamic Treatment Regimes

4.1.1 A historical overview

Perhaps due to the need to identify causal relationships, the study of AIs originated in causal inference with the pioneering works of Robins (see, e.g. Robins, 1986; Robins, 1994, for DTRs). Over an extended period of time, the author introduced three basic approaches for finding effects of time-varying regimes in the presence of confounding variables: the parametric *G-formula* or *G-computation* (Robins, 1986), *structural nested mean models* with the associated method of *G-estimation* (Robins, 1989; Robins, 1992; Robins, 1994), and *marginal structural models* with the associated method of *inverse probability of treatment weighting* (IPW) (Robins, 2000).

A number of methods have subsequently been proposed within statistics, including both frequentist and Bayesian approaches (Lavori & Dawson, 2000; Thall et al., 2000; Thall et al., 2002; Thall et al., 2007). However, all estimate the optimal DTR based on distributional assumptions on the data-generation process via parametric models, and, as such, can easily suffer from model misspecification (Zhao et al., 2015). The first semiparametric method for estimating optimal DTRs was proposed by Murphy (Murphy, 2003), immediately followed by Robins (Robins, 2004), who introduced two alternative approaches using G-estimation. These methods use *approximate dynamic programming*, where 'approximate' refers to the use of an approximation of the value or Q-function introduced in Equation (5), or parts thereof. Thus, they can be considered as the first prototypes of RL-based approaches in the AIs literature.

RL methods represent an alternative approach to estimating DTRs that have gained popularity due to their success in addressing challenging sequential decision-making problems, without the need to fully model the underlying generative distribution. The connection between statistics and RL (previously confined to the computer science and control theory literature) was bridged by Murphy (Murphy, 2005b), who proposed estimating optimal DTRs with Q-learning (Sutton & Barto, 2018; Watkins, 1989). Promptly, a large body of research has embraced the use of Q-learning, integrating various parametric, semiparametric and non-parametric strategies (Chakraborty & Moodie, 2013; Chakraborty & Murphy, 2014; Laber, Linn, & Stefanski, 2014; Murphy, 2005b) to model the Q-function. Q-learning and the semiparametric strategies of Murphy (2003) and Robins (2004) are considered *indirect methods*: optimal DTRs are indirectly obtained by first estimating an optimal objective function (e.g. the Q-function), and then getting the associated (optimal) policy. In contrast, IPW-based strategies (Murphy et al., 2001; Robins, 2000; Wang et al., 2012) seek optimal policies by directly looking for the policy (within a prespecified class of policies) that maximises an objective function (e.g. the expected return), without postulating an outcome model (Zhao et al., 2012); they are regarded as *direct methods*.

In what follows, we review existing RL techniques for developing DTRs focusing on the indirect methods, while an up-to-date review including direct methods can be found in Deliu & Chakraborty (2022). We cover both finite- and indefinite-horizon settings. We emphasise that most of the current work in DTRs deals with finite-horizon problems and *offline learning* procedures that assume access to a collection of observed trajectories. This opposes to the JITAIs tradition—originated with the practical need to deliver AIs in real time—which uses an *online learning* approach for performing data collection and policy optimisation simultaneously. Such procedures can be deployed indefinitely, conditional on practical limitations. In DTRs, the indefinite-horizon setting, particularly suitable for chronic diseases where the number of stages can be arbitrarily large, has been addressed only recently. Nevertheless, it remains relatively understudied.

4.1.2 *Finite-horizon problems*

Finite-horizon DTR problems are designed to identify optimal treatment policies $\boldsymbol{d}^* = \{d_t^*\}_{t=0,\ldots,T}$ over a fixed and known period of time $T < \infty$. Learning methods typically use *offline* RL based on finite (experimental or observational) data trajectories of a sample of say N patients, and causal assumptions about the data (see, e.g. Deliu & Chakraborty, 2022). Each patient trajectory has the form $(X_0, A_0, Y_1, \ldots, X_T, A_T, Y_{T+1})$, with X_0 and $X_1, \ldots, X_T$ the pretreatment and evolving information, respectively, $A_0, \ldots, A_T$ the assigned treatments, and $Y_1, \ldots, Y_{T+1}$ the intermediate outcomes. When a single distal (end-of-study) outcome $\overset{\infty}{Y}$ is considered, all intermediate quantities $\{Y_t\}_{t=1,\ldots,T}$ are taken as 0, and $Y_{T+1} = \overset{\infty}{Y}$, as discussed in Section 2. Note that, especially in observational settings, the decision points t can exhibit greater variability and display less consistent patterns across subjects. In this case, the specific time points t can substantially differ among different individuals, requiring an accurate definition of the admissible time intervals, therefore the N data trajectories, when conducting a DTR analysis.

In finite-horizon problems, RL methods are mainly based on DP or approximate DP procedures. These include Q-learning (Murphy, 2005b), with the Q-function as the objective, and A-learning (Murphy, 2003; Robins, 2004), which focuses on contrasts of conditional mean outcomes. We now discuss the former, assuming throughout this section deterministic policies, that is, policies that map histories $\mathbf{h}$ directly into actions or decisions, that is, $\boldsymbol{d}(\mathbf{h}) = \mathbf{a}$.

Q-learning with function approximation In Section 3, we showed that optimal value functions can be obtained by iteratively solving the Bellman optimality relationship in Equations (10)–(11). In finite-horizon DP problems, this procedure is known as backward induction. However, the iterative process may be memory and computationally intensive, especially for large state and action spaces. Furthermore, traditional DP procedures assume an underlying model for the environment, which is often unknown due to unknown transition probability distributions. Q-learning (Watkins, 1989) offers a powerful and scalable tool to overcome the modelling requirements as well as the computational burden of traditional DP-based methods and constitutes the core of modern RL.

The general idea of Q-learning is that, at each new t, the Q-function is updated based on a previous value and the new acquired information:

$$Q_t^{\boldsymbol{d}}(\mathbf{h_t}, a_t) \leftarrow Q_t^{\boldsymbol{d}}(\mathbf{h_t}, a_t) + \alpha_t \left[Y_{t+1} + \gamma \max_{a_{t+1} \in \mathcal{A}_{t+1}} Q_{t+1}^{\boldsymbol{d}}(\mathbf{h_t} + \mathbf{1}, a_{t+1}) - Q_t^{\boldsymbol{d}}(\mathbf{h_t}, a_t) \right],$$

with α_t a constant that determines to what extent the newly acquired information overrides the old information or how fast learning takes place, and γ a discount factor that balances immediate and future rewards (in finite-horizon problems, it is generally set to one).

The original version of this approach is known as *tabular Q-learning* (Sutton & Barto, 2018). This 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. As the agent selects the actions based on their maximum associated Q-function value, this is equivalent to *exploiting* (recall the notion of exploitation introduced in Section 3.3.2). However, the tabular approach is slow and impractical for large state and action spaces. A powerful and scalable solution to this problem is a more recent version of Q-learning, known as *Q-learning with function approximation* (Murphy, 2005b; Sutton & Barto, 2018). This version first assumes an approximation space for each of the Q-functions in Equation (5), for example, $\mathcal{Q}_t \doteq \left\{Q_t^{\boldsymbol{d}}(\mathbf{h_t}, a_t; \theta_t) : \theta_t \in \Theta_t\right\}$, with parameter space Θ_t a subset of the Euclidean space, and then estimates the optimal stage-t Q-functions Q_t^* backward in time for $t = T, T-1, \ldots, 0$ (Bather, 2000). According to Equation (8), estimating an optimal regime $\widehat{\boldsymbol{d}}^* = \left(d_0^*(x_0), , , , d_1^*(\mathbf{h_1}), \ldots, , , , d_T^*(\mathbf{h_T})\right)$ is equivalent to getting estimates of the optimal Q-functions, or in this case, getting an estimate $\widehat{\theta}_t, t = 1, \ldots, T$, of the parameters, that is,

$$\widehat{d}_t^*(\mathbf{h_t}) = \underset{a_t \in \mathcal{A}_t}{\operatorname{argmax}} \widehat{Q}_t^*(\mathbf{h_t}, a_t) \doteq \underset{a_t \in \mathcal{A}_t}{\operatorname{argmax}} Q_t^*\left(\mathbf{h_t}, a_t; \widehat{\theta}_t\right) \doteq d_t^*\left(\mathbf{h_t}; \widehat{\theta}_t\right).$$

Noticing, for example, that the Q-function is a conditional expectation, we can get the optimal Q-functions as

$$Q_t^*\left(\mathbf{h_t}, a_t; \widehat{\theta}_t\right) \doteq \widehat{\mathbb{E}}_N\left[Y_t + \max_{a_{t+1} \in \mathcal{A}_{t+1}} Q_{t+1}^*\left(\mathbf{h_{t+1}}, a_{t+1}; \widehat{\theta}_{t+1}\right) | \, \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right],$$

with $\widehat{\mathbb{E}}_N$ denoting the empirical mean over a sample of N units. The procedure is illustrated in Supplementary Material E, and a more specific implementation with linear regression is given in Supplementary Material C.

It is important to recognise that the estimated regime $\widehat{\mathbf{d}}^*$ may not be a consistent estimator for the true optimal regime $\boldsymbol{d}^*$, unless all models for the Q-functions are correctly specified. A strategy that may offer robustness to Q-function misspecification is A-learning (Murphy, 2003; Robins, 2004), where 'A' stands for the 'advantage' incurred if the optimal treatment were given as opposed to what was actually given. A-learning represents a class of alternative methods to Q-learning, predicated on the fact that it is not necessary to specify the entire Q-function to estimate an optimal regime. A more in-depth discussion is provided in Supplementary Material D. Schulte et al. (2014) showed that A-learning outperforms Q-learning under misspecifications of Q-function models.

Given that a linear regression model may be quite simple and prone to misspecification, more sophisticated approximators can be used both in Q-learning and in A-learning. These include *support vector regression* (Zhao et al., 2009) and *deep neural networks* (Atan et al., 2018), among others.

Deep Q-network The tremendous success achieved in recent years by RL has been greatly enabled by the use of advanced function approximation techniques such as deep neural networks (DNNs) (Jonsson, 2019; Mnih et al., 2015; Silver et al., 2017), giving rise to the *deep Q-network* algorithm (Mnih et al., 2015). Specifically, at a given time t, a DNN (see Goodfellow et al., 2016, for an overview of existing DNN architectures) is used to fit a model for the

Q-function in a supervised way and then estimate the optimal Q-function: histories $\{\mathbf{H}_{t,i}\}_{i=1,\ldots,N}$ are given as input, and the predicted Q-function values $Q_t^d\left(\mathbf{H}_t, a_t; \widehat{\mathbf{W}}, \widehat{\mathbf{b}}\right)$ associated with each action $a_t \in \mathcal{A}_t$, for example, with $\mathcal{A}_t = \{a_1, \ldots, a_K\}$, are generated as output. $\mathbf{W}$ and $\mathbf{b}$ represent the unknown *weight* and *bias* parameters of a typical DNN; see, for example, the schematic of a *feed-forward neural network* in Figure 5.

Once Q-function estimates are obtained with the DNN, the algorithm proceeds with executing, in an emulator, an action according to an exploration scheme named ϵ-greedy (Sutton & Barto, 2018). This probabilistically chooses between the optimal action so far (i.e. the one with the highest estimated Q-function value) and a random action. Specifically, ϵ is the *exploration* probability for a random action. At the end of the execution sequence, first the Q-function is re-estimated based on the observed reward, and then the DNN parameters are updated using the last Q-function estimates. The pseudo-algorithm is given in Supplementary Material E.

A DNN offers a more flexible and scalable approach, particularly suitable for real-life complexity, high dimensionality and high heterogeneity. Compared with their shallow counterparts, they enable automatic feature representation and can capture complicated relationships (see, e.g. the application in the graft-versus-host disease of Liu et al., 2017). A general limitation of indirect methods such as Q-learning, is that the optimal DTRs are estimated in a two-step procedure: first, the Q-functions are estimated using the data, and then these are optimised to infer the optimal DTR. In the presence of high-dimensional information, even with flexible non-parametric techniques such as DNNs, it is possible that these conditional functions are poorly fitted, with the derived DTR far from optimal. Furthermore, as demonstrated by (Zhao et al., 2012), indirect approaches may not necessarily result in the maximum long-term clinical benefit, motivating direct methods. We refer to (Deliu & Chakraborty, 2022; Tsiatis et al., 2021) for a survey of direct approaches.

Nonetheless, we emphasise that here we present indirect methods in some detail because they are somewhat similar to well-known regression methods that most readers can relate to. Furthermore, many of the methods for developing JITAIs (e.g. Thompson sampling, among the other methods discussed in Section 4.2) are also regression-based methods. Thus, by focusing on regression-type methods across apparently disjoint application domains, we help enhance the synergy between them.

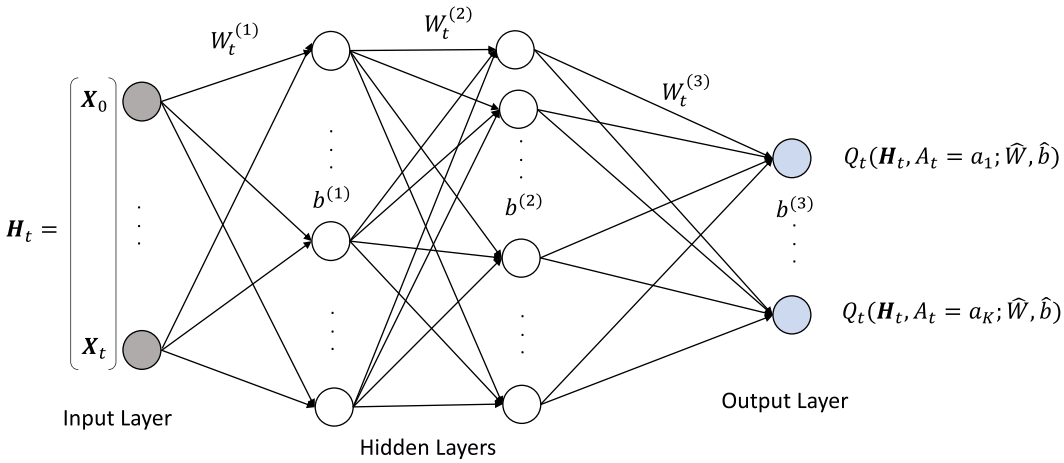

Figure 5. *Schematic of a feed-forward neural network. It is characterised by a set of neurons, structured in four layers ($L = 4$), where each neuron processes the information forward from one layer to the next one. Information is non-linearly transformed according to unknown weights $W^{(l)}$ and bias $b^{(l)}$ parameters, $l = 1, \ldots, L - 1$.*

4.1.3 *Indefinite-horizon problems*

While in computer science there is a vast literature on estimating optimal policies over an increasing time horizon (Sugiyama, 2015; Szepesvari, 2022), that is not the case in DTRs. In fact, by adopting backward induction, most existing methods cannot extrapolate beyond the time horizon in the observed data. Nevertheless, for some chronic conditions or those with very short time steps, including mHealth applications (see Section 2), the time horizon is not definite. Treatment decisions are made continuously throughout the life of a patient, with no fixed time point for the final treatment decision.

To the best of our knowledge, only a limited number of statistical methodologies have been developed for the indefinite-horizon setting. These include the indirect *greedy gradient Q-learning* method of Ertefaie & Strawderman (2018), and the direct *V-learning* approach of Luckett et al. (2020), who proposed to search for an optimal policy over a prespecified class of policies. More recently, a minimax framework called proximal temporal consistency learning was proposed (Zhou et al., 2022). We now detail the first two approaches, while for the third, we refer the reader to the original work in Zhou et al. (2022).

Greedy gradient Q-learning The first extension to indefinite-time horizons in DTRs was proposed in Ertefaie & Strawderman (2018), under the time-homogeneous Markov assumption (see Section 3.3.1). Although not imposed by general DTR methods, such assumption overcomes the need for backward induction, and exemplifies inference by working with time-independent Q-functions.

We adopt the notation of the previous sections and introduce an absorbing state c, representing a loss-to-follow-up, for example, death, event. We assume that at each time t, covariates X_t take values in a finite state space $\mathcal{X}^{\boxtimes} \doteq \mathcal{X} \cup \{c\}$, with $\mathcal{X} \cap \{c\} = \varnothing$. Let the action space $\mathcal{A}_x$ be finite and defined by covariate information such that $\mathcal{A}_x$ consists of $0 < K_x \leq K$ treatments, with K being the total number of treatments over the time horizon. For any t such that $X_t = c$, let $A_x = \mathcal{A}_c = \{u\}$, where u stands for 'undefined'. Now, denoting a stopping time (e.g. death) by $\widetilde{T} \doteq \inf\{t > 0 : X_t = c\}$, individual trajectories are of the form $\left(X_0, A_0, R_1, \ldots, X_{\widetilde{T}-1}, A_{\widetilde{T}-1}, R_{\widetilde{T}}, X_{\widetilde{T}}\right)$. Note that $\mathbb{P}\left(\widetilde{T} < \infty | X_0, A_0\right) = 1$, regardless of (X_0, A_0). Based on these specifications, the indefinite time-t Q-function for regime $\boldsymbol{d}(\mathbf{h_t}) = \boldsymbol{d}(x_t) = \boldsymbol{d}(x)$, for $x \in \mathcal{X}$, is given by:

$$Q^{\boldsymbol{d}}(x, a) \doteq \mathbb{E}_{\boldsymbol{d}}[\mathbf{R_t} \,|\, X_t = x_t, A_t = a_t] = \mathbb{E}_{\boldsymbol{d}}\left[\sum_{\tau=0}^{\infty} \gamma^{\tau - t} Y_{\tau+1} \middle| X_t = x_t, A_t = a_t\right].$$

We set $Q^*(c, a) = 0$ because the return is 0 after an individual is lost to follow-up.

For estimating an optimal DTR, Q-learning is proposed. Let $Q(x, a; \theta^*)$ be a parametric model for $Q^*(x, a)$ indexed by $\theta^* \in \Theta \subseteq \mathbb{R}^n$, with $n \geq 1$, and postulate a linear model with interactions, that is, $Q(x, a; \theta^*) = \theta^{*T} f(x, a)$, with $f(x, a)$ being a known feature vector summarising the state and treatment pair. To ensure $Q^*(c, a) = 0$, we also need $f(c, a) = 0$. Now, defining $f(X_t, A_t) \doteq \nabla_{\theta^*} Q(X_t, A_t; \theta^*)$, with ∇ the gradient, Bellman optimality suggests and motivates the following unbiased estimating function for θ^*:

$$\widehat{D}(\theta^*) = \widehat{\mathbb{P}}_N \left\{ \sum_{t=0}^{T-1} \left(Y_{t+1} + \gamma \max_{a \in \mathcal{A}_{X_{t+1}}} Q(X_{t+1}, a; \theta^*) - Q(X_t, A_t; \theta^*) \right) f(X_t, A_t) \right\}. \quad (14)$$

Note that the estimating function in Equation (14) is a non-convex and non-differentiable function of θ^*, which complicates the estimation process. Under regularity conditions, the

authors suggested that any solution $\widehat{\theta}^*$ can be equivalently defined as a minimiser of $\widehat{M}(\theta^*) \doteq \widehat{D}(\theta^*)^T \widehat{S}^{-1} \widehat{D}(\theta^*)$, with $\widehat{S} \doteq \widehat{\mathbb{P}}_N \left\{ \sum_{t=0}^{T-1} f(X_t, A_t)^{\otimes 2} \right\}$, and $x^{\otimes 2} \doteq xx^T$, for any vector x. If $\widehat{\theta}^* = \arg\min_{\theta^* \in \Theta} \widehat{M}(\theta^*)$ is the unique solution, then $\widehat{Q}^*(x, a) = Q\left(x, a; \widehat{\theta}^*\right)$, and the corresponding optimal regime is given by $\widehat{\boldsymbol{d}}^* = \arg\max_{a \in \mathcal{A}_x} Q\left(x, a; \widehat{\theta}^*\right)$.

V-learning The greedy gradient Q-learning approach based on Equation (14) involves a non-smooth max operator that makes estimation difficult without large amounts of data (Laber, Linn, & Stefanski, 2014; Linn et al., 2017). Motivated by an mHealth application, where policy estimation is continuously updated in real time as data accumulate (starting with small sample sizes), an alternative method is proposed in Luckett et al. (2020). Under the same time-homogeneous MDP assumption, provided that interchange of the sum and integration is justified, the authors consider the value function

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

and follow a direct approach to directly maximise estimated values over a prespecified class of policies. In light of the Bellman equation in Equation (9), it follows that, for any function f defined on the state space $\mathcal{X}_t$, the following importance-weighted variant is satisfied:

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

Let $V^{\boldsymbol{d}}(x; \theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^n$, be a model for $V^{\boldsymbol{d}}(x)$. Assume that $V^{\boldsymbol{d}}(x; \theta)$ is differentiable everywhere in θ for fixed x and $\boldsymbol{d}$. Then, the proposed estimating equation is given by

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

Again, $\widehat{\theta}$ can be obtained by minimising $\widehat{M}(\theta) \doteq \widehat{\Lambda}(\theta)^T \widehat{S}^{-1} \widehat{\Lambda}(\theta) + \lambda \mathcal{P}(\theta)$, with $\widehat{S}$ a positive definite matrix in $\mathbb{R}^{n \times n}$, λ a tuning parameter, and $\mathcal{P}: \mathbb{R}^n \to \mathbb{R}_+$ a penalty function. The estimated optimal regime $\widehat{d}^*$ is the argmax of $V^{\boldsymbol{d}}\left(x; \widehat{\theta}\right)$. Compared with greedy gradient Q-learning, V-learning requires modelling both the policy and the value function, but not the data-generating process. In addition, by directly maximising the estimated value over a class of policies (see Luckett et al., 2020, for more details), it overcomes the issues of the non-smooth max operator in Equation (14). The method is applicable over indefinite horizons and is suitable for both offline and online learning, which is typical in JITAIs.

4.2 Just-in-time Adaptive Interventions in Mobile Health

Unlike DTRs, where the number of decision points is generally small, JITAIs are defined upon a random and indefinitely large number of times. They are carried out in dynamic environments with the scope of capturing rapid changes in an individual user's context and needs (Nahum-Shani et al., 2015; Nahum-Shani et al., 2018). Methodologies for optimising JITAIs require the ability to learn nearly continuously, with no definite time horizon. Furthermore, learning is performed *online* as data accumulate, often using 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 optimise; that is, $\boldsymbol{\pi} = \boldsymbol{d}$. Thus, existing methods for DTRs, which mainly target a finite-time horizon problem and are implemented *offline* (e.g. Q-learning), are not directly applicable to JITAIs. Furthermore, by carrying over an entire history of an individual, they may not be feasible from a computational perspective.

As discussed in Section 3, the standard approach for developing JITAIs is given by contextual MABs (Tewari & Murphy, 2017), an intermediate solution between MABs (Auer et al., 2002) and the full-RL approach used in DTRs. With a few exceptions, contextual MAB algorithms applied in mHealth rely on two fundamental bandit strategies, originally implemented in advertising: the *linear upper confidence bound* (LinUCB) (Li et al., 2010) and the *linear Thompson sampling* (LinTS) (Agrawal & Goyal, 2013).

4.2.1 Contextual bandits with upper confidence bound exploration

So far, optimal AIs have typically been identified by finding the optimal Q-functions recursively with the Bellman relationships. In contrast, LinUCB (Chu et al., 2011; Li et al., 2010) employs the underlying idea of MABs, where the optimal policy is the set of the optimal stage-t arms, for all t, defined individually for each time t as: $d_t^* \doteq A_t^* \doteq \arg\max_{a_t \in \mathcal{A}} \mathbb{E}(Y_{t+1}|, |X_t = x_t|, , A_t = a_t)$. Notice that this objective function represents a myopic version of the Q-function in Equation (5) by taking $\gamma = 0$ and reflects a stochastic MAB setting where contexts are IID and one can ignore the previous history given the last state x_t (see also Equation (12)), that is,

$$Q_t^{\boldsymbol{\pi}}(\mathbf{h_t}, a_t) = Q_t^{\boldsymbol{\pi}}(x_t, a_t) = \mathbb{E}[Y_{t+1} \,|\, X_t = x_t,\, A_t = a_t], \quad \forall a_t \in \mathcal{A}_t, \quad \forall x_t \in \mathcal{X}_t, \quad \forall t \in \mathbb{N}.$$

The specific solution of LinUCB is based on performing an efficient exploration by favouring arms for which a confident value has not been estimated yet and avoiding arms which have shown a low reward with high confidence. This confidence is measured by the *upper confidence bound* of the expected reward value for each arm. The underlying assumption is that the conditional expected reward is a linear function of a context-action feature f, that is, $\mathbb{E}[Y_{t+1}|X_t, A_t] = f(X_t, A_t)^T\mu$, with $\mu \in \mathbb{R}^n$ the unknown coefficient vector associated with the feature f. In this work, we consider general features f (constructed via linear basis, polynomials or splines expansion, among others; see, e.g. Marsh & Cormier, 2002) rather than a standard linear function that may fail to capture non-linearities in the data.

Under the linear model assumption, the LinUCB idea is to estimate at each time t an upper bound, say $U_t(a_t)$, for the expected reward of each arm a_t. The LinUCB estimator is defined as

$$\widehat{U}_t(a_t) \doteq \widehat{Q}_t^{\boldsymbol{\pi}}(x_t, a_t) + \alpha s_t(a_t) = f(X_t = x_t,\, A_t = a_t)^T \widehat{\mu}_t + \alpha s_t(a_t), \tag{15}$$

where $\alpha > 0$ is a tuning parameter that controls the trade-off between exploration and exploitation: small values of α favour exploitation while larger values of α favour exploration. The first part $f(X_t, , A_t = a_t)^T\widehat{\mu}_t$, with $\widehat{\mu}_t \doteq B_t^{-1}b_t$ being an estimate of μ_t, reflects the current point estimate of the expected reward of the arm a_t. The second term represents the confidence we have in this estimate, resembling a typical confidence interval: $s_t(a_t) \doteq \sqrt{f(x_t, a_t)^T B_t^{-1} f(x_t, a_t)}$ reflects the uncertainty, or the standard deviation, and α can be viewed as a generalisation of the critical value. Note also that B_t^{-1} and b_t are analogous to the terms '$\left(X^TX\right)^{-1}$' and 'X^TY', respectively, appearing in the ordinary least squares estimator for a standard linear regression model with $\mathbb{E}[Y|X] = X^T\mu$. If we assume a ridge penalised estimation strategy, with penalty parameter $\lambda \geq 0$, these values are recursively computed at each time t taking into account previously

explored arms: $B_t \doteq \lambda \mathbb{I}_n + \sum_{\tau=0}^{t-1} f(x_\tau, \widetilde{a}_\tau)^T f(x_\tau, \widetilde{a}_\tau)$ and $b_t \doteq \sum_{\tau=0}^{t-1} f(x_\tau, \widetilde{a}_\tau)^T Y(x_\tau, \widetilde{a}_\tau)$, where $\{\widetilde{a}_\tau \doteq \arg\max_{a_\tau \in \mathcal{A}} U_\tau(a_\tau)\}_{\tau=0,1,\ldots,t-1}$ are the optimal arms estimated at previous times and $\mathbb{I}_n$ is the identity matrix of order n. A schematic of the LinUCB approach is provided in Supplementary Material E.

Several variations of LinUCB were proposed in the bandit literature. These include (i) the *linear associative RL* strategy (Auer, 2003), based on singular value decomposition rather than ridge regression; (ii) generalised linear models, aiming to accommodate more complex models either for the reward (Filippi et al., 2010; Li et al., 2017) or the environment (Urteaga & Wiggins, 2019); (iii) non-parametric modelling of the reward function, such as Gaussian processes (Srinivas et al., 2012); and (iv) a neural network-based feature construction which overcomes the linear reward assumption (Zhou et al., 2020). More recently, in addition to the (bandit) optimisation goal, attention has been given to statistical objectives. To illustrate, in a similar context as ours, that is, behavioural science, Dimakopoulou et al. (2019) introduced balancing methods from the causal inference literature. Specifically, to make the algorithm less prone to bias, authors proposed to weight each observation with the estimated inverse probability of a context being observed for an arm. This algorithm helps to reduce bias, particularly in misspecified cases, at the cost of increased variance.

Successful applications of LinUCB in mHealth can be found in Forman et al. (2019) and Paredes et al. (2014). The former developed a LinUCB-based intervention recommender system for delivering stress management strategies (upon user's request in a mobile app), with the goal of maximising stress reduction. After 4 weeks of study, participants who received LinUCB-based recommendations demonstrated to use more constructive coping behaviours. Similarly, in Forman et al. (2019), a pilot study was conducted to evaluate the feasibility and acceptability of an RL-based behavioural weight loss intervention system. Participants were randomised between a non-optimised group, an individually-optimised group (individual reward maximisation), and a group-optimised (group reward maximisation) group. The study showed that the LinUCB-based optimised groups have strong promise in terms of the outcome of interest, not only being feasible and acceptable for participants and coaches, but also achieving desirable results at roughly one-third the cost.

4.2.2 *Contextual bandits with Thompson sampling exploration*

Although Thompson sampling (TS) (Thompson, 1933) has been introduced more than 80 years ago, it has only recently reemerged as a powerful tool for online decision-making, due to its optimal empirical and theoretical properties. Under the same linear reward assumption as in LinUCB, Agrawal and Goyal (Agrawal & Goyal, 2013) proposed a generalisation of TS to a contextual setting. Rooted in a Bayesian framework, the idea of TS is to select arms according to their posterior probability of being optimal, that is, by maximising the posterior reward distribution, at each time t. The policy π at each time t is thus explicitly defined as:

$$\begin{aligned}\pi_t(a) &= \mathbb{P}\big(Q_t^{\pi}(x_t, a) \geq Q_t^{\pi}(x_t, a'),\ \forall a' \neq a \mathbf{H_t} = \mathbf{h_t}\big) \\ &= \mathbb{P}(\mathbb{E}[Y_{t+1} \,|\, X_t = x_t, A_t = a] \geq \mathbb{E}[Y_{t+1} \,|\, X_t = x_t, A_t = a'],\ \forall a' \neq a \mathbf{H_t} = \mathbf{h_t}), \quad t = 0, 1, \ldots,\end{aligned} \tag{16}$$

where the conditioning term $\mathbf{H_t} = \mathbf{h_t}$ reflects the posterior nature of this probability and should not be confused with the conditioning terms of the Q-function. The TS policy has been shown to be asymptotically optimal, meaning that it matches the asymptotic lower bound of the regret introduced by Lai and Robbins (Lai & Robbins, 1985).

The typical way to implement TS is iterative and involves a posterior sampling procedure (see, e.g. Chapelle & Li, 2011). For example, in the common case of a Gaussian reward model

with variance v^2, that is, $Y_t|\mu, f(X_t, A_t) \sim \mathcal{N}\left(f(X_t, A_t)^T\mu, v^2\right)$, considering a Gaussian prior for the regression coefficients vector μ, for example, $\mu \sim \mathcal{N}(\mathbf{0}_n, , \sigma^2\mathbb{I}_n)$, at each time t, the optimal arm is the one that maximises the posterior estimated expected reward, or $f(X_t, A_t)^T\widetilde{\mu}_t$. The posterior nature is reflected in $\widetilde{\mu}_t$, which represents a sample from the estimated posterior distribution, given by $\mathcal{N}\left(\widehat{\mu}_t, , v^2B_t^{-1}\right)$; here $\widehat{\mu}_t \doteq B_t^{-1}b_t$ is the posterior mean, with B_t and b_t defined in the same way as for LinUCB. The iterative LinTS procedure is given in Supplementary Material E.

Given the history up to time t and $f(X_t, A_t)$, the LinUCB allocation policy is deterministic, in the sense that the t-step arm or intervention a_t is uniquely determined as the one that maximises the upper confidence bound in Equation (15); all the other arms have a null probability of being assigned. In contrast, LinTS can be regarded as a randomised scheme, where each of the admissible arms has a positive probability of being assigned to an individual, independent of their history. In other words, given the history up to time t and $f(X_t, A_t)$, the LinTS allocation policy, as defined in Equation (16), is still random. In terms of exploration, LinUCB allows exploration through the uncertainty term $s_t(a_t)$, while LinTS achieves it through the random draws from the posterior distribution, or, equivalently, through the probability $\pi_t(a_t)$ in Equation (16). Note that the standard deviation of LinUCB and LinTS is of the same order. In fact, in LinTS $Y_t|\mu_t, f(x_t, a_t) \sim \mathcal{N}\left(f(x_t, a_t)^T\widehat{\mu}_t, \ v^2f(x_t, \ a_t)^TB_t^{-1}f(x_t, \ a_t)\right)$, and by definition $f(x_t, a_t)^TB_t^{-1}f(x_t, a_t) = s_t(a_t)$.

Similarly to LinUCB, many extensions have been considered. In the mHealth literature, specifically addressing complex likelihood functions, Eckles and Kaptein (2019) formulated a Bootstrap TS version to replace the posterior by an online bootstrap distribution of the point estimate $\widehat{\mu}_t$ at each time t. The approach offers improved robustness to model misspecifications (due to the robustness of the bootstrap approach), and it can be easily adapted to dependent observations, a common feature of behavioural sciences. Tackling a different issue, namely, sparse and noisy data, Tomkins et al. (2021) introduced *Intelligent Pooling*, a generalised version of LinTS with a Gaussian mixed-effects linear model for the reward. By explicitly modelling heterogeneity between individuals and within an individual over time, the method demonstrates a better promise of personalisation, even in small groups of users.

Action-centered Thompson sampling Motivated by potential non-stationarities in mHealth problems, Greenewald et al. (2017) generalised the stationary linear model of LinTS to a non-stationary and non-linear version, where the expected reward model is formalised as

$$\mathbb{E}(Y_{t+1}|X_t = x_t, A_t = a_t) = f(x_t, a_t)^T\mu I(a_t \neq 0) + g_t(x_t), \quad t \in \mathbb{N}, \tag{17}$$

with $g_t(x_t)$ being the main non-stationary component that can vary based on past information, but not on current action, and $I(\cdot)$ denoting the indicator function. Compared with LinTS, here the reward is conceived as a combination of a baseline reward (associated with a 'do nothing' or control arm, say $a_t = 0$), which is entirely determined by $g_t(x_t)$, and a treatment or action effect (associated with non-control arms, say $a_t \neq 0$). The latter is a linear function of the context-action feature $f(x_t, a_t) \in \mathbb{R}^n$ with $\mu \in \mathbb{R}^n$ the unknown parameters (resembling the fixed component characterising LinTS), but consider contexts X_t chosen by an adversary based on the history up to time t. The term adversarial in contextual MABs can refer to the context and the reward generation mechanism: when both contexts and rewards are allowed to be chosen arbitrarily by an adversary, no assumptions on the generating process are made (see also Section 3.3.2), and data can be non-IID.

To avoid user habituation, caused, for example, by the delivery of too many interventions, and to prevent the algorithm from converging to an ineffective deterministic policy, a stochastic chance constraint on the size of the probabilities of delivering the non-control arm is considered. That is, given two fixed probability thresholds $\pi_{\min}$ and $\pi_{\max}$, with $0 < \pi_{\min} < \pi_{\max} < 1$, the probability of assigning a non-control arm is given by

$$\pi_t(a) = \max\Big(\pi_{\min}, \min\Big(\pi_{\max}, \mathbb{P}\Big(f(x_t, a)^T \mu > 0\Big)\Big)\Big), \quad a \neq 0,$$

where $\mathbb{P}\Big(f(x_t, a)^T \mu > 0\Big)$ represents the expected treatment effect of a non-control arm a, while μ reflects the parameter posterior distribution as in LinTS. The proposed strategy, named *action-centered TS*, can be viewed as a hierarchical two-step procedure, where the first step involves estimating the optimal non-control arm, that is, the one that maximises the expected treatment effect or reward as in classical LinTS, and the second step is to randomly select between a control and non-control arm $A_t \neq 0$. To allow for better comparability with LinTS, both algorithms are described with their pseudo-code in Supplementary Material E.

The specific use of the term '*action-centered*' reflects the estimation procedure for the unknown parameters μ: Due to the arbitrarily complex baseline reward $g_t(x_t)$, the authors propose to work with the differential reward, defined as $Y_{t+1}(X_t, A_t) - Y_{t+1}(X_t, 0) = f(X_t, A_t)^T \mu I(A_t > 0) + \varepsilon_{t+1}$, which has the scope to eliminate the component $g_t(x_t)$; this allows the derivation of an unbiased estimator, and we refer to (Greenewald et al., 2017) for further details. The authors also showed that the action-centered TS achieves performance guarantees similar to LinTS, while allowing for non-linearities in the baseline reward. Additional theoretical improvements are given in Kim & Paik (2019) and Krishnamurthy et al. (2018). Here, a relaxation of the action-independent assumption of the component $g_t(x_t)$ in Equation (17) is considered, making the reward model entirely non-parametric.

From a practical viewpoint, the algorithm has been empirically evaluated in the *HeartSteps* study (Klasnja et al., 2015; Liao et al., 2020), an mHealth physical activity study of great interest both in biostatistics and in the RL/bandit literature. In this context, Liao at el. (Liao et al., 2020), for example, incorporated into the differential reward model an 'availability' variable, stating whether the user is available to receive an intervention or not.

4.3 Insights on Current Methodological Differences and Their Drivers

So far, a broad literature documented and demonstrated the premise of RL in both types of AIs. However, the practical methodological realities of the two remain apparently disjoint or with little commonalities. Why does Q-learning, a popular algorithm for estimating DTRs, have no practical use in JITAIs, where simplified frameworks are used? Is it reasonable to adopt simplified RL formulations given the nature of mHealth applications? Can we expect Thompson sampling, largely employed in JITAIs, to dictate the next generation of DTRs? Or ultimately, should we expect a convergence, dictated, for example, by a greater synergy between the two areas, or should we regard them as unrelated?

Although these questions were partially covered throughout the previous sections, here we offer a systematic synthesis of the main differences and their drivers. We propose a number of insights that may guide the thinking about the future of the two areas and their potential relationship (or lack thereof), convergence, or complementarity. In Section 5, case studies supporting this discussion will be illustrated.

4.3.1 *Offline versus online learning: a different primary objective*

One of the main differences between the DTR and JITAI settings is in the role optimising schemes such as RL have from the data acquisition to the data analysis. In DTRs, RL is *typically* (we will mention a few exceptions shortly) implemented in an offline manner: we assume data have already been collected, and RL serves for estimating an optimal regime from a batch of N IID data trajectories. RL does not guide the treatment decisions as new data arise, and thus has no role in the data collection process. Data are typically generated based on routine clinical practice (leading to observational data such as EHRs), or according to a randomised study where interventions are assigned with, for example, fixed and equal probability at each stage. On the contrary, in JITAIs, RL is *often* the main determinant of data collection: its scope is to determine and deliver, based on accumulating data, the right interventions in real time so as to benefit the most of the study participants. Clearly, while this process will not directly affect the sample, it will affect the intervention assignment and thus the final data.

This difference is mainly driven by the typical primary objective in each AI area, especially in the case of experimental studies, where an intervention decision can be made within the study. For DTR studies, the purpose of a potential experimental trial is generally to evaluate and compare treatments; only in subsequent analyses/phases, the goal embraces identifying and eventually assigning (to a future population) an optimal estimated regime. Nonetheless, it is worth noting that data arising from experimental studies, in particular SMARTs, are still limited. In fact, due to cost and complexity in the design and implementation, SMARTs are relatively few in number. In this landscape, the availability of observational data has guided a broad literature on DTR, leaving space for offline learning only. Yet most of the literature has focused on illustrating the application of statistical methodology, rather than informing clinical practice (Mahar et al., 2021). One of the main challenges for an adequate translation into practice is given by the difficulty in verifying the necessary conditions for causal inference, in addition to the high heterogeneity of both patient populations and treatment implementation, which requires accurate pre-analysis considerations. We refer to section 15.3.1.1 of Deliu & Chakraborty (2022) for two relevant examples on constructing DTRs with observational data.

Unlike DTRs, in JITAIs, delivering optimal AIs during the course of a program or the trial in order to optimise users' experience and engagement with the mHealth device remains the primary goal. Clearly, this trend is facilitated by the use of mobile devices and by the type of intervention (more behavioural and less clinical). Nonetheless, an increasing amount of population is using mobile devices for behavioural health support outside of experimental contexts in their everyday life. This anticipates a large amount of observational-type of data for estimating JITAIs, along with several challenges for analyses, spanning from the large portion of missing data to the high heterogeneity in terms of users' behaviour in using the mHealth tool (e.g. number and distance between decision points or availability to interact with the device).

There are some exceptions, and these may indicate a potential convergence. For example, in the context of infectious diseases, the application of a fixed randomisation strategy for a prolonged period is neither ethical nor feasible. To this end, the use of TS has been proposed to learn and assign an optimal treatment strategy online (Laber et al., 2018). Notably, the MAB choice in this setting addresses some challenges that cannot be directly addressed with offline methods such as Q-learning, including: (i) scarsity of data at the onset of an epidemic, (ii) high dimensionality and scalability with respect to state and action spaces, and (iii) a long and indefinite time horizon. Similarly, there have been theoretical works (see, e.g. Cheung et al., 2015; Wang et al., 2022) that tried to incorporate online adaptations within a SMART to skew the randomisation probabilities towards the most promising treatments.

4.3.2 Simplifying assumptions and domain aspects

Delivery of JITAIs in mHealth is carried out primarily through the simplified RL framework of contextual MABs. This simplification is essentially dictated by the strong assumption that actions A_t have a momentary effect on rewards Y_{t+1}, but do not affect the distribution of the next states $\{X_\tau\}_{\tau \geq t+1}$. Essentially, one sets the discount parameter γ to 0, and looks for the optimal in-the-moment action. Domain knowledge envisions that such an assumption is reasonable in many mHealth applications, where information such as weather, time of the day and GPS location, among others, is momentary, as is also its effect. However, one may certainly question the validity of such a choice or whether we should use a larger γ, full- or MDP-RL, or other strategies. In clinical conditions, this is often unrealistic: the effect of a treatment may be observed at different times and may be affected by delayed or carryover effects. In mHealth, the study of phenomena such as habituation and delayed rewards is becoming increasingly common. Incorporating these considerations, in addition to allowing for non-stationarities (as in the action-centered TS (Greenewald et al., 2017)), may favour the RL methods used in DTRs and advance knowledge discovery.

4.3.3 Learning efficiency

Clearly, solving a full-RL or an MDP-RL problem is much more computationally demanding than solving a contextual MAB problem. The discount rate γ is strongly related to computational expense: the larger is the γ, or the farther we look ahead, the higher is the computational burden. By choosing small values of γ, one trades off the optimality of the learned policy for computational efficiency, which is a critical aspect in high-dimensional problems. Notably, contextual data in many mHealth applications is highly private. For this reason, much of the computation has to be done locally on mobile devices, with the risk of severely impacting battery life.

4.3.4 Inference and real-time inference

A key aspect that has been extensively studied in DTRs is the problem of inference (see Deliu & Chakraborty, 2022; Tsiatis et al., 2021, for a recent overview). This aspect has been neglected in JITAIs, where the primary goal is oriented towards reward optimisation, or alternatively, participants' benefit. Learning about intervention regimes and drawing generalised conclusions is often beyond the scope of their delivery. However, even when the focus is on the ongoing study itself, how can we support the development of high-quality JITAIs without adequately assessing the effectiveness of the sequence of interventions delivered by the mobile device? Clearly, compared with (mostly behavioural) JITAIs, delivering DTRs involves a higher risk, as each intervention (often a drug) can have a substantial impact on patients' lives. Thus, among other critical points, including cost, this ethical aspect has long limited the learning and delivery of DTRs online.

We emphasise that adaptive data-collection settings driven by RL present major challenges for statistical inference due to potential strong imbalances in arms allocations and the underlying sequential nature of the data. The problem is nowadays well documented (Deliu et al., 2021; Hadad et al., 2021), and recent solutions have borrowed tools from causal inference (Zhang et al., 2021).

While this discussion is motivated by the apparent differences between the DTR and JITAI methods, we have shown that there are exceptions, and that the use of a specific RL method should be driven by the applied problem at hand and its peculiarities (population, disease or condition, underlying domain characteristics and ethical concerns). We acknowledge that computational costs (memory and time), as well as technological limitations, still play a dominant role in JITAIs; whereas in DTRs, the main drivers relate to ethical aspects and the costs associated with running high-quality intervention studies such as SMARTs.

Going beyond methodology, we conclude this section by suggesting that, despite the fact that DTRs and JITAIs originated and developed within two different domains while following a similar—if not the same—goal, they could often have a complementary role. In fact, if construction of an optimised DTR is part of the objectives of an experimental study (even if secondary), JITAIs could be utilised to deliver behavioural interventions to enhance both adherence to treatment and engagement with the health study, all in support of high-quality data collection, with limited deviations from study protocol.

4.4 Further Considerations and Future Directions

We notice that our focus in this work has been mainly devoted to the development of optimised AIs and, more specifically, to the use of RL to solve this dynamic optimisation problem. Several other aspects are crucial for rigorously, validly and ethically operating in the space of AIs.

First, an adequate framework for causal inference is necessary. In this work, we only mentioned it in passing (see Section 3.1) and assumed that this foundational block and the underlying assumptions (such as unconfounding) hold. Although RL practitioners often consider problems in which the data are unconfounded, healthcare practitioners often need to make inferences using offline RL from observational data, where unconfouness is a major concern. Thus, many RL methods may not be directly applicable in practice, and learning optimal DTRs/JITAIs should account for the fact that observed actions might be affected by unobserved confounders. To address this issue, a growing literature is studying confounding bias and possible corrections under a *partially observed Markov decision process* (POMDP) (Bennett & Kallus, 2023; Miao et al., 2018; Uehara et al., 2023). This states that there exist unobserved (hidden or latent) state variables, say $Z_t \in \mathcal{Z}_t$, such that (X_t, Z_t, Y_t) forms an MDP in the sense outlined in Section 3.3.1, with $t \in \mathbb{N}$; that is, $(X_{t+1}, Z_{t+1}, Y_{t+1}) \perp \{(X_\tau, Z_\tau, Y_\tau)\}_{\tau=0}^{t-1} \mid (X_t, Z_t, A_t)$, leading to a trajectory distribution:

$$P_{\boldsymbol{\pi}}^{\mathrm{POMDP}} \doteq p_0(x_0, z_0) \prod_{t \geq 0} \pi_t(a_t|x_t, z_t) p_{t+1}\left(x_{t+1}, z_{t+1}, y_{t+1}|x_t, z_t, a_t\right).$$

Note that a POMDP allows some state variables Z_t to remain unobserved by the agent; therefore, it is a weaker assumption compared with a Markovian structure. Furthermore, the unobserved state can be regarded as a source of unobserved confounding. Within the POMDP framework, and drawing upon prior research on causal identification utilising proxy variables (Miao et al., 2018), Bennett and colleagues (Bennett & Kallus, 2023) introduced the so-called *proximal reinforcement learning*. This method aims to address confounding by employing two independent proxies for confounders, one assumed to be conditionally independent from treatments given confounders, and the other assumed to be independent from outcomes given treatment and confounders. This strategy conforms to the proximal causal inference literature, which accommodates unmeasured confounding within specific causal structures. Interested readers can explore further discussions in the works of Miao et al. (2018) and Zivich et al. (2023).

Second, if on one side the increasing technological and computational sophistication has led to new biomedical data sources (e.g. data from mobile devices and EHRs) and new algorithmic solutions (e.g. DNN or RL), on the other side it has posed some unique new challenges to be inclusively addressed. We refer to (Zicari, 2013) for a comprehensive survey. Among these, ethical and societal concerns about fairness, accountability and transparency are becoming increasingly relevant (see, e.g. the cross-disciplinary view adopted by top ML conferences such as the *ACM Conference on Fairness, Accountability, and Transparency*). In the context of ML and RL, fairness considerations are gaining a certain depth when it comes to decision-making in healthcare and medical research (see, e.g. (Chen et al., 2023; Chien et al., 2022; Mitchell

et al., 2021)). In particular, as optimal AIs are estimated by optimising an average cumulative outcome, due to individuals' diversity in their responsiveness to treatment and adverse effects, the estimated optimal AI may be suboptimal, risky, or even detrimental to certain underrepresented or disadvantaged subpopulations. Efforts to address this issue have been presented in Fang et al. (2023), Li et al. (2023) and Zhu et al. (2024) for both single-stage and dynamic treatment regimes. The typical approach is to work with a fairness or risk-aware optimisation problem, where a constraint is placed on the tail performance (fairness; Fang et al., 2023) or on some unwanted side effect (risk; Li et al., 2023; Zhu et al., 2024). A detailed discussion on safe/risk-sensitive RL in healthcare is provided in Appendix A of Li et al. (2023). Let interest be in a single final outcome $\overset{\infty}{Y}$, and denote by $\mathscr{L}_r^{\boldsymbol{d}}\left(\overset{\infty}{Y}\right) \doteq \inf\{y : F(y) \geq r\}$ the r-th quantile of $\overset{\infty}{Y}$ under policy $\boldsymbol{d}$, with F and $r \in (0, 1)$ denoting the cumulative distribution function and a quantile level of interest, respectively. Focusing on fairness, (Fang et al., 2023) proposed looking at the following fairness-oriented optimisation problem for the population mean $\mathbb{E}_{\boldsymbol{d}}\left(\overset{\infty}{Y}\right)$:

$$\max_{\boldsymbol{d}} \mathbb{E}_{\boldsymbol{d}}\left(\overset{\infty}{Y}\right), \quad \text{subject to} \quad \mathscr{L}_r^{\boldsymbol{d}}\left(\overset{\infty}{Y}\right) \geq q,$$

where $q \in \mathbb{R}$ is a predefined threshold guarantee on the tail performance. Estimation of the expectation and quantile can be more or less complex depending on whether one considers a single or a sequence of decision rules, and we refer to the original work in Fang et al. (2023) for theoretical analyses in both cases.

Future work may also complement this project by covering and integrating from other healthcare domains that use RL. A non-exhaustive list of examples is given in Deliu (2021) and Yu et al. (2023), which includes, among others, the design of *adaptive clinical trials* (U.S. Department of Health and Human Services Food and Drug Administration, 2019). In such settings, by utilising and processing accumulated data in an online fashion, RL and MAB methods could contribute to making clinical trials more flexible, efficient, informative and ethical (Pallmann et al., 2018; Villar et al., 2015). Fairness has also been the subject of recent debates (Chien et al., 2022). All of these aspects may deserve a dedicated space, and we aim to pursue this research direction as a separate piece of work in the near future.

5 Reinforcement Learning in Real Life: Case Studies

This section complements the methodological framework introduced so far with its real-world implementation. Guided by two case studies we conducted in the space of DTRs and mHealth, respectively, we: (i) illustrate the applicability of RL as well as the main challenges researchers face in applying these methods in practice; and (ii) provide a concrete illustration of the main divergence between RL methods in the two areas. We start with a brief introduction of the two studies, before summarising and comparing their main characteristics side by side in Table 3.

5.1 Dynamic treatment regimes: PROJECT QUIT – FOREVER FREE

Based on a two-stage SMART design, this study aimed to develop/compare internet-based behavioural interventions for smoking cessation and for relapse prevention. The primary objective, interesting the first stage, known as *PROJECT QUIT*, only, was to find an optimal multifactor behavioural intervention to help adult smokers quit smoking (see Strecher et al., 2008, for details). The second stage, known as *FOREVER FREE*, was a follow-on study designed to: (i) help *PROJECT QUIT* participants who quit smoking stay non-smoking and (ii) offer a second chance to those who failed to give up smoking at the previous stage. These two stages

Table 3. *Summary of two case studies, in DTRs and JITAIs in mHealth, respectively, that used RL*

	PROJECT QUIT – FOREVER FREE	*DIAMANTE*
Study design	*SMART*	*MRT*
Primary objective	To find an optimal internet-based behavioural intervention for smoking cessation and relapse prevention (based on the first stage of the SMART only)	To develop and evaluate the effectiveness of a JITAI solution for enhancing physical activity, by means of an RL-based text-messaging system
Secondary objective	To find an optimal DTR (based on the entire two-stage SMART design)	To assess the effectiveness of the JITAI solution on the distal outcome (i.e. depression)
Role of RL in the design and analysis	Secondary, used *offline* for secondary post-data collection analysis	Primary, used *online* in the design (data-collection phase) during interim analyses
Number, frequency and distance between decision points	Two decision points at a minimum distance of 6 months: one at the first-stage entry (the 6-month-long *PROJECT QUIT*) and one after completion of the first stage at the second-stage entry (the 6-month-long *FOREVER FREE*)	Daily, with around 180 decision points over a 6-month-long study: intervention decisions are made at a random time interval (Factor T in Figure 3) and can distance from 14 to 22 h.
Model choice: interventions and tailoring variables	A parsimonious model with the statistically significant elements of the primary regression analyses: • Stage-1 model included two intervention factors (each at two levels) and three covariates; • Stage-2 model included two intervention arms, the three stage-1 covariates and an additional covariate represented by the intermediate outcome (quit status at the end of stage 1); • Interactions between interventions and covariates were included as well	A high-dimensional model including • All baseline variables shown to be relevant in the literature and other time-varying covariates; • An action space given by the combinations of the $4 \times 5 \times 4$ factor levels; • Action-action and action-contextual interactions were also included
Choice of the RL strategy for optimising interventions	Offline learning based on • Q-learning with a linear model, chosen for its simplicity and interpretability; • A *soft-thresholding* estimator (within the Q-learning framework) to address the vexing problem of non-regularity	Online learning based on • The computationally efficient and randomised TS algorithm to mitigate bias and to enable causal inference (Rosenberger et al., 2019); • Self-regularisation (implemented within TS) to deal with the high dimensionality and avoid overfitting; • An initial uniform random 'burn-in' period or, more appropriately, an 'internal pilot' to acquire some prior data to feed into the main algorithm and speed up learning
Primary outcome, that is, the reward variable directly targeted by the intervention	A final distal outcome related to smoking cessation and defined as the seven-day point prevalence of smoking (i.e. whether or not the participant smoked even a single cigarette in the last 7 days prior to the end of the study stages)	A proximal outcome related to physical activity and defined as the steps change from 1 day to another, starting the steps count from the time an intervention message is sent
Handling of missing data in the reward variable	• Descriptive checks, revealing a more or less uniform dropout across the different intervention arms, and • Complete case analysis (Chakraborty et al., 2010), as well as sensitivity analysis	• Online imputation with the *last observation carried forward*, and • Multiple imputation as a sensitivity analysis, To provide reliable final estimates and avoid harmful impacts (due to technical errors in

(*Continues*)

Table 3 *(Continued)*

	PROJECT QUIT – FOREVER FREE	*DIAMANTE*
Study design	*SMART*	*MRT*
	of multiply-imputed data (Chakraborty, 2009), To avoid sub-optimal policies due to potentially different patterns across different interventions	collecting observations) on online decision making
Other study challenges	Inference, high-dimensionality, feature extraction, sample size considerations and power analysis (see also Deliu & Chakraborty, 2022; Laber, Lizotte, et al., 2014)	Inference, non-stationarity and delayed reward, model misspecification and noisy data, users' disengagement, sample size considerations and power analysis (see also Figueroa et al., 2021; Liao et al., 2020)

were then considered together with the goal of finding an optimal DTR over the entire SMART study period; this was a secondary objective of the study. RL was not used in the design phase; in other words, this is not an instance of *online learning*. The RL-type learning happened *offline* on completion of the data collection. Detailed results from this secondary analysis can be found in Chakraborty (2009) and Chakraborty et al. (2010).

5.2 Just-in-time Adaptive Inteventions in Mobile Health: DIAMANTE

Based on an MRT design (illustrated earlier in Section 2), the primary objective of the trial was to evaluate the effectiveness of an RL-based text-messaging system for delivering JITAIs to encourage individuals to become more physically active. In this case, RL was implemented *online*, with interventions continuously optimised according to users' time-varying individual data. To evaluate the optimised JITAI solution, users were assigned to different study groups (see Figure 3), including a static (non-optimised) group and the experimental RL-based adaptive group. For further details, we refer to Aguilera et al. (2020) and Figueroa et al. (2021).

6 Conclusions

In this work, under a unified framework that brings together DTRs and JITAIs in mHealth under the area of adaptive interventions, we showed how these problems can be formalised as RL problems. With a sincere hope to enhance synergy between the methodological and applied communities, we provided a comprehensive state-of-the-art survey on RL strategies for AIs, augmenting the methodological framework with real examples and challenges. Then, we discussed the main methodological divergences in the two AI domains.

Notably, while the two areas are ideally sharing the same problem of finding optimal policies (in line with the RL framework), their priorities are not always aligned due to historical links or domain restrictions. DTRs are mainly focused on offline estimation and identification of causal nexuses, while JITAIs are mainly engaged in online regret performances, neglecting the problem of inference. Only recently, a small body of literature started to examine the possibility of inferential goals in JITAIs, questioning the validity of traditional statistics in adaptively-collected data (Deliu et al., 2021; Hadad et al., 2021; Zhang et al., 2021). The ML community has led the way in addressing such issues, often borrowing tools from causal inference. For example, the 'stabilising policy' approach of Zhang et al. (2021) is analogous to the 'stabilised weights' of the causal inference literature (Robins, 2000). Similarly, the

adaptively-weighted IPW estimator in Hadad et al. (2021) is inspired by the IPW estimator in Robins (2000). Furthermore, an increased attention is paid to real-time or online inference to evaluate the effectiveness of JITAIs online (see, e.g. Dimakopoulou et al., 2019; Dimakopoulou et al., 2021).

Despite the insufficiently mature field of mHealth, with a relatively small number of methodological studies for a rigorous evaluation of RL methods for JITAIs, their popularity in real life has grown remarkably (see Supplementary Material A). In contrast, in DTRs, the use of RL has been extensively evaluated in theoretical works, but its application in the real world is still very limited. Most existing DTR studies use real data only as motivational or illustrative examples. The few clinical studies focus mainly on offline learning based on observational data (e.g. EHRs) and deep learning methodologies, which limits interpretability. The explanatory drivers may be related to (1) the lack of existing guidelines for developing optimal, yet statistically valid, DTRs; (2) the clinical setting itself, characterised by high costs, ethical concerns and inherent complexities, which makes experimentation hard; (3) the lack of definition of AI components and the RL dynamics for the specific disease. When defining the reward function, for instance, one may need to account for multiple objectives and the presence of unstructured data, among other prior knowledge. Even from an implementation perspective, while several software packages exist for DTRs, these are often suitable only under simplified settings, for example, continuous and positive rewards. We recognise that the area of mHealth, mostly related to behavioural aspects rather than clinical, may have fewer concerns in terms of treatment costs and risks.

To the best of our knowledge, this represents the first piece of work that bridges the domains of DTRs and JITAIs under a unique umbrella intersecting RL and AIs. Our hope is that such a unified common ground, where different methodological and applied disciplines can easily cooperate, would help unlock the potential of exploring the opportunity RL offers in AIs and benefiting from it in a statistically justifiable way. For example, by using the rich resources on inference made available by the DTR literature, the JITAI literature may extend its goal beyond within-trial optimisation. Similarly, if SMARTs were to be used in practice more often, in addition to collecting high-quality experimental data, decisions could also be optimised online, benefiting trial participants as well (see, e.g. Cheung et al., 2015; Laber et al., 2018), as done in JITAIs.

We also hope that our contribution may incentivise greater synergy and cooperation between the statistical and ML communities to support applied domains in the conduct of high-quality real-world studies. We recognise that this cooperation is very timely to support both the development of real-world DTR studies and to assist the spread of mHealth applications with reliable and reproducible workflows.

ACKNOWLEDGMENTS

The authors would like to thank the two anonymous reviewers and Eric Laber for the constructive feedback received. Nina Deliu acknowledges support from the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) and from Sapienza University (Finanziamenti di ateneo per la ricerca scientifica, Avvio alla Ricerca Tipo 1: AR11916B8913234D). Joseph J. Williams was supported by the Office of Naval Research (N00014-21-1-2576) and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2019-06968). Bibhas Chakraborty would like to acknowledge support from the start-up funding from the Duke-NUS Medical School, as well as the Academic Research Fund Tier 2 grant (MOE T2EP20122-0013) from the Ministry of Education, Singapore.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

Agrawal, S. & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13*, pp. 1220–1228.

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., Jay Williams, J. & 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.

Almirall, D., Nahum-Shani, I., Sherwood, N.E. & Murphy, S.A. (2014). Introduction to SMART designs for the development of adaptive interventions: With application to weight loss research. *Transl. Behav. Med.*, **4**(3), 260–274.

Atan, O., Jordon, J. & van der Schaar, M. (2018). Deep-Treat: Learning optimal personalized treatments from observational data using neural networks. *Proc. AAAI Conf. Artif. Intell.*, **32**(1).

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, **3**, 397–422.

Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R.E. (2002). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, **32**(1), 48–77.

Bather, J. (2000). *Decision Theory: An Introduction to Dynamic Programming and Sequential Decisions*. Wiley Interscience Series in Systems and Optimization. New York: Wiley, Chichester.

Beam, A.L. & Kohane, I.S. (2018). Big data and machine learning in health care. *JAMA*, **319**(13), 1317.

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

Bennett, A. & Kallus, N. (2023). Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Oper. Res.*, page opre.2021.0781.

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. New York: Springer.

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

Chakraborty, B., Murphy, S. & Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.*, **19**(3), 317–343.

Chakraborty, B. & Murphy, S.A. (2014). Dynamic treatment regimes. *Ann. Rev. Stat. Appl.*, **1**(1), 447–464.

Chapelle, O. & Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Chen, R.J., Wang, J.J., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S. & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.*, **7**(6), 719–742.

Cheung, Y.K., Chakraborty, B. & Davidson, K.W. (2015). Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program: SMART with adaptive randomization. *Biometrics*, **71**(2), 450–459.

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

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

Collins, F.S. & Varmus, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.*, **372**(9), 793–795.

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

Collins, L.M., Murphy, S.A. & Bierman, K.L. (2004). A conceptual framework for adaptive preventive interventions. *Prev. Sci.*, **5**(3), 185–196.
Dawson, R. & Lavori, P. W. (2012). Efficient design and inference for multistage randomized trials of individualized treatment policies. *Biostatistics*, **13**(1), 142–152.
Deliu, N. (2021). *Reinforcement Learning in Modern Biostatistics: Benefits, Challenges and New Proposals*. PhD Thesis, University of Rome La Sapienza, Rome.
Deliu, N. (2023). Reinforcement learning for sequential decision making in population research. *Qual. Quant.*
Deliu, N. & Chakraborty, B. (2022). Dynamic treatment regimes for optimizing healthcare. In *The Elements of Joint Learning and Optimization in Operations Management*, Eds. X. Chen, S. Jasin, & C. Shi, Springer Series in Supply Chain Management pp. 391–444. Cham: Springer International Publishing.
Deliu, N., Williams, J.J. & Villar, S.S. (2021). Efficient Inference Without Trading-off Regret in Bandits: An Allocation Probability Test for Thompson Sampling. arXiv:2111.00137 [cs, stat].
Deo, R.C. (2015). Machine learning in medicine. *Circulation*, **132**(20), 1920–1930.
Dimakopoulou, M., Ren, Z. & Zhou, Z. (2021). Online multi-armed bandits with adaptive inference. In *Advances in Neural Information Processing Systems*, volume **34**, pp. 1939–1951. Curran Associates, Inc.
Dimakopoulou, M., Zhou, Z., Athey, S. & Imbens, G. (2019). Balanced linear contextual bandits. *In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI'19*, pp. 3445–3453, Honolulu, Hawaii, USA: AAAI Press.
Eckles, D. & Kaptein, M. (2019). Bootstrap Thompson sampling and sequential decision problems in the behavioral sciences. *SAGE Open*, **9**(2).
Ertefaie, A. & Strawderman, R.L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, **105**(4), 963–977.
Fang, E.X., Wang, Z. & Wang, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *J. Am. Stat. Assoc.*, **118**(543), 1733–1746.
Fernández-Loría, C. & Provost, F. (2022). Causal decision making and causal effect estimation are not the same … and why it matters. *INFORMS J. Data Sci.*, **1**(1), 4–16.
Figueroa, C.A., Aguilera, A., Chakraborty, B., Modiri, A., Aggarwal, J., Deliu, N., Sarkar, U., Jay Williams, J. & Lyles, C.R. (2021). Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. *J. Am. Med. Inform. Assoc.*, **28**(6), 1225–1234.
Figueroa, C.A., Deliu, N., Chakraborty, B., Modiri, A., Xu, J., Aggarwal, J., Jay Williams, J., Lyles, C. & Aguilera, A. (2022). Daily motivational text messages to promote physical activity in university students: Results from a microrandomized trial. *Ann. Behav. Med.*, **56**(2), 212–218.
Filippi, S., Cappe, O., Garivier, A. & Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume **23**. Curran Associates, Inc.
Forman, E.M., Kerrigan, S.G., Butryn, M.L., Juarascio, A.S., Manasse, S.M., Ontañón, S., Dallal, D.H., Crochiere, R. J. & Moskow, D. (2019). Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *J. Behav. Med.*, **42**(2), 276–290.
Garnett, C., Crane, D., West, R., Brown, J. & Michie, S. (2019). The development of *Drink Less*: An alcohol reduction smartphone app for excessive drinkers. *Transl. Behav. Med.*, **9**(2), 296–307.
Goldberg, Y. & Kosorok, M.R. (2012). Q-learning with censored data. *Ann. Stat.*, **40**(1).
Goldstein, S.P., Evans, B.C., Flack, D., Juarascio, A., Manasse, S., Zhang, F. & 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. *Int. J. Behav. Med.*, **24**(5), 673–682.
Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning. Adaptive Computation and Machine Learning*. Cambridge, Massachusetts: The MIT Press.
Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F. & Celi, L.A. (2019). Guidelines for reinforcement learning in healthcare. *Nat. Med.*, **25**(1), 16–18.
Greenewald, K., Tewari, A., Klasnja, P. & Murphy, S. (2017). Action centered contextual bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pp. 5979–5987, Red Hook, NY, USA: Curran Associates Inc.
Hadad, V., Hirshberg, D.A., Zhan, R., Wager, S. & Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proc. Natl. Acad. Sci.*, **118**(15), e2014602118.
Hardeman, W., Houghton, J., Lane, K., Jones, A. & Naughton, F. (2019). A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. *Int. J. Behav. Nutr. Phys. Act.*, **16**(1), 31.
Hernan, M.A. & Robins, J.M. (2023). *Causal Inference: What If*. Boca Raton: CRC Press.
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)*. New York, N.Y: Springer.
Jonsson, A. (2019). Deep reinforcement learning in medicine. *Kidney Dis.*, **5**(1), 18–22.

Kasy, M. & Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, **89**(1), 113–132.

Kim, G.-S. & Paik, M.C. (2019). Contextual multi-armed bandit algorithm for semiparametric reward model. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3389–3397. PMLR.

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

Kosorok, M.R. & Laber, E.B. (2019). Precision medicine. *Ann. Rev. Stat. Appl.*, **6**(1), 263–286.

Krishnamurthy, A., Wu, Z.S. & Syrgkanis, V. (2018). Semiparametric contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2776–2785. PMLR.

Kumar, H., Li, T., Shi, J., Musabirov, I., Kornfield, R., Meyerhoff, J., Bhattacharjee, A., Karr, C., Nguyen, T., Mohr, D., Rafferty, A., Villar, S., Deliu, N. & Williams, J.J. (2024). Using adaptive bandit experiments to increase and investigate engagement in mental health. *Proc. AAAI Conf. Artif. Intell.*, **38**(21), 22906–22912.

Kumar, S., Murphy, S.A., and Rehg, J.M., editors (2017). *Mobile Health: Sensors, Analytic Methods, and Applications*. Springer International Publishing: Imprint: Springer, Cham, 1 edition.

Kumar, S., Nilsen, W.J., Abernethy, A., Atienza, A., Patrick, K., Pavel, M., Riley, W. T., Shar, A., Spring, B., Spruijt-Metz, D., Hedeker, D., Honavar, V., Kravitz, R., Craig Lefebvre, R., Mohr, D.C., Murphy, S.A., Quinn, C., Shusterman, V. & Swendeman, D. (2013). Mobile health technology evaluation. *Am. J. Prev. Med.*, **45**(2), 228–236.

Laber, E.B., Linn, K.A. & Stefanski, L.A. (2014). Interactive model building for Q-learning. *Biometrika*, **101**(4), 831–847.

Laber, E.B., Lizotte, D.J., Qian, M., Pelham, W.E. & Murphy, S.A. (2014). Dynamic treatment regimes: technical challenges and applications. *Electron. J. Stat.*, **8**(1).

Laber, E.B., Meyer, N.J., Reich, B.J., Pacifici, K., Collazo, J.A. & Drake, J.M. (2018). Optimal treatment allocations in space and time for on-line control of an emerging infectious disease *J. R. Stat. Soc. Series C, Appl. Stat.*, **67**(4), 743–770.

Lai, T. & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, **6**(1), 4–22.

Lattimore, T. & Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press, 1 edition.

Lavori, P.W. & Dawson, R. (2000). A design for testing clinical strategies: Biased adaptive within-subject randomization. *J. R. Stat. Soc. A. Stat. Soc.*, **163**(1), 29–38.

Lavori, P.W. & Dawson, R. (2004). Dynamic treatment regimes: Practical design considerations. *Clin. Trials*, **1**(1), 9–20.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D. & Murphy, S. (2012). A 'SMART' design for building individualized treatment sequences. *Annu. Rev. Clin. Psychol.*, **8**(1), 21–48.

Li, L., Chu, W., Langford, J. & Schapire, R.E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, Raleigh North Carolina USA. ACM.

Li, L., Lu, Y. & Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, pp. 2071–2080, Sydney, NSW, Australia. JMLR.org.

Li, Y., Zhou, W. & Zhu, R. (2023). Quasi-optimal Reinforcement Learning with Continuous Actions. In The Eleventh International Conference on Learning Representations.

Liao, P., Greenewald, K., Klasnja, P. & Murphy, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, **4**(1), 1–22.

Linn, K. A., Laber, E.B. & Stefanski, L.A. (2017). Interactive Q-learning for quantiles. *J. Am. Stat. Assoc.*, **112**(518), 638–649.

Liu, X., Deliu, N. & Chakraborty, B. (2023). Microrandomized trials: Developing just-in-time adaptive interventions for better public health. *Am. J. Public Health*, **113**(1), 60–69.

Liu, Y., Logan, B., Liu, N., Xu, Z., Tang, J. & Wang, Y. (2017). Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 380–385, Park City, UT. IEEE.

Luckett, D.J., Laber, E. B., Kahkoska, A.R., Maahs, D.M., Mayer-Davis, E. & Kosorok, M.R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *J. Am. Stat. Assoc.*, **115**(530), 692–706.

Lunceford, J.K., Davidian, M. & Tsiatis, A.A. (2002). Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, **58**(1), 48–57.

MacKinnon, D.P., Fairchild, A.J. & Fritz, M.S. (2007). Mediation analysis. *Annu. Rev. Psychol.*, **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 Med. Res. Methodol.*, **21**(1), 39.

Marsh, L. & Cormier, D. (2002). *Spline Regression Models*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America.
Miao, W., Geng, Z. & Tchetgen Tchetgen, E. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, **105**(4), 987–993.
Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Ann. Rev. Stat. Appl.*, **8**(1), 141–163.
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. & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, **518**(7540), 529–533.
Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018). *Foundations of Machine Learning. Adaptive Computation and Machine Learning*. The MIT Press, Cambridge, Massachusetts, second edition edition.
Murphy, S.A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Series B Stat. Methodology*, **65**(2), 331–355.
Murphy, S.A. (2005a). An experimental design for the development of adaptive treatment strategies. *Stat. Med.*, **24**(10), 1455–1481.
Murphy, S.A. (2005b). A generalization error for Q-learning. *J. Mach. Learn. Res.*, **6**, 1073–1097.
Murphy, S.A., Collins, L. & Rush, A.J. (2007). Customizing treatment to the patient: Adaptive treatment strategies. *Drug Alcohol Depend.*, **88**, S1–S3.
Murphy, S.A., Lynch, K.G., Oslin, D., McKay, J.R. & TenHave, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug Alcohol Depend.*, **88**, S24–S30.
Murphy, S.A., Van Der Laan, M.J., Robins, J.M., and Conduct Problems Prevention Research Group (2001). Marginal mean models for dynamic regimes. *J. Am. Stat. Assoc.*, **96**(456), 1410–1423.
Nahum-Shani, I. & 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. & Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychol.*, **34**(Suppl), 1209–1219.
Nahum-Shani, I., Smith, S.N., Spring, B.J., Collins, L.M., Witkiewitz, K., Tewari, A. & Murphy, S.A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Ann. Behav. Med.*, **52**(6), 446–462.
Naughton, F. (2017). Delivering 'just-in-time' smoking cessation support via mobile phones: Current knowledge and future directions. *Nicotine Tob. Res.*, **19**(3), 379–383.
Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.*, **5**(4).
Oyebode, O., Fowles, J., Steeves, D. & Orji, R. (2022). Machine learning techniques in adaptive and personalized systems for health and wellness. *Int. J. Human–Comput. Interact.*, 1–25.
Pallmann, P., Bedding, A.W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L.V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M.R., Villar, S.S., Wason, J.M.S., Weir, C.J., Wheeler, G.M., Yap, C. & Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med.*, **16**(1), 29.
Paredes, P., Gilad-Bachrach, R., Czerwinski, M., Roseway, A., Rowan, K. & Hernandez, J. (2014). PopTherapy: coping with stress through pop-culture. In Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth'14, pp. 109–117, Brussels, BEL. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.
Pelham, W.E., Hoza, B., Pillow, D.R., Gnagy, E.M., Kipp, H.L., Greiner, A.R., Waschbusch, D.A., Trane, S.T., Greenhouse, J., Wolfson, L. & Fitzpatrick, E. (2002). Effects of methyphenidate and expectancy on children with ADHD: Behavior, academic performance, and attributions in a summer treatment program and regular classroom settings. *J. Consult. Clin. Psychol.*, **70**(2), 320–335.
Pfammatter, A.F., Nahum-Shani, I., DeZelar, M., Scanlan, L., McFadden, H.G., Siddique, J., Hedeker, D. & Spring, B. (2019). SMART: Study protocol for a sequential multiple assignment randomized controlled trial to optimize weight loss management. *Contemp. Clin. Trials*, **82**, 36–45.
Pike, A.C. & Robinson, O.J. (2022). Reinforcement learning in patients with mood and anxiety disorders vs control individuals: A systematic review and meta-analysis. *JAMA Psychiatry*, **79**(4), 313.
Puterman, M.L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics section. New York: Wiley.
Rajkomar, A., Dean, J. & Kohane, I. (2019). Machine learning in medicine. *N. Engl. J. Med.*, **380**(14), 1347–1358.
Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.*, **58**(5), 527–535.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Modell.*, **7**(9–12), 1393–1512.
Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, **79**(2), 321–334.
Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, Eds. L. Sechrest, H. Freeman, & A. Mulley, pages 113–159. NCHSR, U.S. Public Health Service.
Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Commun. Stat. - Theory Methods*, **23**(8), 2379–2412.
Robins, J.M. (2000). Marginal Structural Models versus Structural nested Models as Tools for Causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials, The IMA Volumes in Mathematics and Its Applications*, Eds. M.E. Halloran & D. Berry, pp. 95–133, New York, NY: Springer.
Robins, J.M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data, Lecture Notes in Statistics*, Eds. D.Y. Lin and P.J. Heagerty, pp. 189–326. New York, NY: Springer.
Rosenberger, W.F., Uschner, D. & Wang, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Stat. Med.*, **38**(1), 1–12.
Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**(5), 688–701.
Schulte, P.J., Tsiatis, A.A., Laber, E.B. & Davidian, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Stat. Sci.: Rev. J. Inst. Math. Stat.*, **29**(4), 640–661.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, **550**(7676), 354–359.
Srinivas, N., Krause, A., Kakade, S.M. & Seeger, M.W. (2012). Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory*, **58**(5), 3250–3265.
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. & Pomerleau, O.F. (2008). Web-based smoking-cessation programs. *Am. J. Prev. Med.*, **34**(5), 373–381.
Sugiyama, M. (2015). *Statistical Reinforcement Learning: Modern Machine Learning Approaches*. Chapman and Hall/CRC.
Sutton, R.S. & Barto, A.G. (2018). Reinforcement learning: An introduction. *Adaptive Computation and Machine Learning series*. The MIT Press, Cambridge, Massachusetts, 2 edition.
Szepesvari, C. (2022). *Algorithms for Reinforcement Learning*. Springer.
Tewari, A. & Murphy, S.A. (2017). From Ads to Interventions: Contextual Bandits in Mobile Health. In *Mobile Health*, Eds. J.M. Rehg, S.A. Murphy, and S. Kumar, pp. 495–517. Cham: Springer International Publishing.
Thall, P.F., Millikan, R.E. & Sung, H.-G. (2000). Evaluating multiple treatment courses in clinical trials. *Stat. Med.*, **19**(8), 1011–1028.
Thall, P.F., Sung, H.-G. & Estey, E.H. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *J. Am. Stat. Assoc.*, **97**(457), 29–39.
Thall, P.F., Wooten, L.H., Logothetis, C.J., Millikan, R.E. & Tannir, N.M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Stat. Med.*, **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.
Tomkins, S., Liao, P., Klasnja, P. & Murphy, S. (2021). IntelligentPooling: Practical Thompson sampling for mHealth. *Mach. Learn.*, **110**(9), 2685–2727.
Tsiatis, A.A., Davidian, M., Holloway, S.T. & Laber, E.B. (2021). *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman & Hall/CRC, Boca Raton. OCLC: 1259526921.
U.S. Department of Health and Human Services Food and Drug Administration (2019). *Adaptive Design Clinical Trials for Drugs and Biologics: Guidance for Industry.* Technical report, Washington DC, USA: US Department of Health and Human Services Food and Drug Administration, CDER, CBER.
Uehara, M., Kiyohara, H., Bennett, A., Chernozhukov, V., Jiang, N., Kallus, N., Shi, C. & Sun, W. (2023). Future-Dependent Value-Based Off-Policy Evaluation in POMDPs. In Thirty-seventh Conference on Neural Information Processing Systems.
Urteaga, I. & Wiggins, C.H. (2019). (Sequential) Importance Sampling Bandits. arXiv:1808.02933 [cs, stat].
Van Otterlo, M. & Wiering, M. (2012). Reinforcement learning and Markov decision processes. In *Reinforcement Learning*, Eds. M. Wiering and M. van Otterlo, volume **12**, pp. 3–42. Berlin Heidelberg, Berlin, Heidelberg: Springer.

Villar, S.S., Bowden, J. & Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Stat. Sci.*, **30**(2).
Voils, C.I., Chang, Y., Crandell, J., Leeman, J., Sandelowski, M. & Maciejewski, M.L. (2012). Informing the dosing of interventions in randomized trials. *Contemp. Clin. Trials*, **33**(6), 1225–1230.
Wahed, A.S. & Tsiatis, A.A. (2006). Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika*, **93**(1), 163–177.
Wang, J., Wu, L. & Wahed, A.S. (2022). Adaptive randomization in a two-stage sequential multiple assignment randomized trial. *Biostatistics*, **23**(4), 1182–1199.
Wang, L., Rotnitzky, A., Lin, X., Millikan, R.E. & Thall, P.F. (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *J. Am. Stat. Assoc.*, **107**(498), 493–508.
Watkins, C.J.C.H. (1989). *Learning from delayed rewards*. PhD, Cambridge, UK: King's College, Cambridge University.
Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M. & Hochberg, I. (2017). Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *J. Med. Internet Res.*, **19**(10), e338.
Yu, C., Liu, J., Nemati, S. & Yin, G. (2023). Reinforcement learning in healthcare: a survey. *ACM Comput Surv*, **55**(1), 1–36.
Zhang, J. & Bareinboim, E. (2020). Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11012–11022. PMLR.
Zhang, K.W., Janson, L. & Murphy, S.A. (2021). Statistical Inference with M-estimators on adaptively collected data. *Adv. Neural Inf. Process. Syst.*, **34**, 7460–7471.
Zhao, Y., Kosorok, M.R. & Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Stat. Med.*, **28**(26), 3294–3315.
Zhao, Y., Zeng, D., Rush, A. J. & Kosorok, M.R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.*, **107**(499), 1106–1118.
Zhao, Y.-Q., Zeng, D., Laber, E.B. & Kosorok, M.R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Am. Stat. Assoc.*, **110**(510), 583–598.
Zhou, D., Li, L. & Gu, Q. (2020). Neural contextual bandits with UCB-based exploration. *In Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pp. 11492–11502.
Zhou, W., Zhu, R. & Qu, A. (2022). Estimating optimal infinite horizon dynamic treatment regimes via pT-learning. *J. Am. Stat. Assoc.*, 1–14.
Zhu, S., Shen, W., Fu, H. & Qu, A. (2024). Risk-aware restricted outcome learning for individualized treatment regimes of schizophrenia. *Ann. Appl. Stat.*, **18**(2).
Zhu, W., Zeng, D. & Song, R. (2019). Proper inference for value function in high-dimensional Q-learning for dynamic treatment regimes. *J. Am. Stat. Assoc.*, **114**(527), 1404–1417.
Zicari, R.V. (2013). Big data: Challenges and Opportunities. In Akerkar, R., editor, *Big Data Computing*, pp. 245–269. Chapman and Hall/CRC.
Zivich, P.N., Cole, S.R., Edwards, J.K., Mulholland, G.E., Shook-Sa, B.E. & Tchetgen Tchetgen, E.J. (2023). Iintroducing proximal causal inference for epidemiologists. *Am. J. Epidemiol.*, **192**(7), 1224–1227.

[*Received January 2024; accepted May 2024*]