SAPIENZA
UNIVERSITÀ DI ROMA

Reinforcement Learning in Modern Biostatistics: Benefits, Challenges and New Proposals

Ph.D. Programme *School of Statistical Sciences*

Ph.D in *Methodological Statistics* – XXXIII Cycle

Candidate

Nina Deliu

ID number 1494917

Thesis Advisor

Prof. Pierpaolo Brutti

Co-Advisors

Prof. Bibhas Chakraborty

Prof. Joseph Jay Williams

Thesis defended on May 24, 2021
in front of a Board of Examiners composed by:

Prof. Alessandra Luati (chair)

Prof. Claudio Agostinelli

Prof. Bruno Scarpa

Reinforcement Learning in Modern Biostatistics: Benefits, Challenges and New Proposals

Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: April 11, 2021

Author's email: nina.deliu@uniroma1.it

To the greatest source of energy
I have even seen,
My Mom

Alla più grande fonte di energia
che abbia mai visto,
Mia Mamma

Acknowledgments

"Life is not found in atoms or molecules or genes as such, but in organization; not in symbiosis, but in synthesis."
- Edwin Grant Conklin

If I could give a "romantic" subtitle to this thesis, I would probably opt for something like "The Partial Voyage of a Statistician in the Bright- and Dark-Side of the Machine Learning World of Promises in Healthcare". Each word has a specific meaning, which can be understood by reading the pages of this thesis, but their enthusiastic intersection and cooperation represents the summary of my Ph.D. adventure: three years spent in a multidisciplinary environment, working separately and simultaneously with statisticians, data scientists, computer scientists, engineers, phychologists, and clinicians. I have to admit that it was not the easiest experience of my life, as jointly integrating and accepting all the different perspectives and objectives is not a straightforward arrival; but it was definitely worth the journey, providing a great input for improvements in communicability, knowledge sharing and science advancing, with a 360*-degree view and analysis approach.*

Thus, my first thanks go to the multidisciplinary team who welcomed me, showed me what a team is, and thought me how forces are joined together with the common objective of advancing knowledge discovery. Thanks to Bibhas, my main mentor during this years, thanks to whom everything had a start (from the other side of the world, Singapore); to Joseph, an agglomeration of enthusiasm and energy, who represents the central pole of the team (pole, because he's almost at the North Pole, in Canada); to Sofia, currently my highest source of inspiration, I cannot express my gratitude to you for your humanity and your ability to share with generosity your knowledge. Thanks to all the other people who joined their forces in this team (Audrey, Anna, the IAI Lab and the Berkeley's group); and a sincere gratitude goes to the two reviewers (Audrey Durand and Ying Kuen Cheung) who revised and contributed with their time, dedication and precious feedback to enhance the content of this work. Last but not least, thanks to my supervisor Pierpaolo, who adviced and supported me since the very beginning of my university life.

My deepest and sincere thanks go to my family, for their continuous and unparalleled love and help. I am grateful to my sisters and my brother-in-law Andrea for always encouraging me and being there as friends. I am forever indebted to my parents for giving me the opportunities and experiences that made me who I am.

My friends, I now realize that without you my Ph.D adventure, my everyday adventures, would have never been as special as they were, almost as special as you are, and as special as my thanks for you are. Suzan, Camilla, Alessia, Veronica, Giulia, Riccardo, Alessia, Mohammed: you mean so much to me!

Finally, I want to share my heartfelt gratitude to life. Which brought me out to meet you all, and recently the most unexpectedly perfect person I could have ever imagined, Marco: "Sei il mio sole nel cuore".

P.S. Of course, thanks to coffee; and thanks to midnight hours, who gave me inspiration for putting in words the efforts of my Ph.D years.

Abstract

Applications of reinforcement learning (RL) for supporting, managing and improving decision-making are becoming increasingly popular in a variety of medicine and healthcare domains where the problem has a sequential nature. By continuously interacting with the underlying environment, RL techniques are able to learn by trial-and-error on how to take better actions in order to maximize an outcome of interest over time. However, if on one hand RL offers a new powerful framework, on the other hand it poses some unique challenges for data analysis and interpretability, which call for new statistical techniques in both predictive and descriptive learning.

Notably, several methodological challenges, for which the contribution of the biostatistical community may play a crucial role, limit the use of RL in real life. In an aim to bridge the statistics and RL communities, we start by assimilating the different existing RL terminologies, notations and approaches into a coherent body of work, and by translating them from a machine learning (ML) to a statistical perspective. Then, through a comprehensive methodological review, we report and discuss the state-of-the-art RL-based research in healthcare. Two main applied domains emerged: 1) adaptive interventions (AIs), encompassing both dynamic treatment regimes and just-in-time adaptive interventions in mobile health (mHealth); and 2) adaptive designs of clinical trials, specifically dose-finding designs and adaptive randomization. We illustrate existing RL-based methods in these areas, discussing their benefits and existing open problems that may impact their application in real life.

A major barrier to adopting RL in real-world experiments is the lack of clarity on how statistical analyses and inference are impacted. In clinical trials for example, if on one side, to achieve the practical (and more ethical) goal of improving patients' benefits, RL may have better abilities in terms of maximising clinical outcomes by adaptively randomizing participants to the best evidence-based treatment; on the other side, to achieve the scientific goal of e.g., discovering whether one treatment is more effective compared to a control treatment, less is known about their inferential properties. Through a simulation study, we investigate the challenges of conducting hypothesis testing from data collected through a class of RL, i.e., multi-armed bandits (MABs), outlining the harms MAB algorithms can cause to traditional statistical tests' type-I error and power. This empirical evaluation provides guidance to two alternative ways of pursuing improved statistical hypothesis testing: 1) to explore ways of modifying the test statistic using knowledge of the adaptive data collection nature; 2) to modify the algorithm or framework for a more sensitive problem to both statistical inference as well as reward maximization. Focusing on the Thompson Sampling (a randomized MAB strategy), we show how a modified version of it results in an optimal intermediate between these two objectives.

These findings can provide insights into how challenges can be surmounted by bridging machine learning, statistics, and applied sciences, to conduct adaptive experiments in the real-world, aiming to simultaneously help individuals and advance scientific research. We finally combine our methodological knowledge with a motivating mHealth study for improving physical activity, to illustrate the tremendous collaboration opportunities between statistics and RL researchers in the space of developing adaptive interventions into the increasingly growing area of mHealth.

Contents

Chapter 1

General Introduction

"What matters is not the enclosure of the work within a harmonious figure, but the centrifugal force produced by it - a plurality of language as a guarantee of a truth that is not merely partial."
- Italo Calvino, *Six Memos For The Next Millennium*

In the era of big data and digital innovation, healthcare is undergoing a process of rapid and dramatic change, with clinical decision support systems acquiring a vital role for both developing and improving care delivery (Sultan, 2015; Jiang *et al.*, 2017). Increasing technological sophistication has lead to new biomedical data sources, such as electronic health records (EHRs) and mobile/wearable devices, that are generating exponentially more data, with an increasing complexity (Dash *et al.*, 2019; Toga *et al.*, 2015; He *et al.*, 2019). If on one side, this flood of data offers a new powerful resource for improving quality and costs of healthcare and for advancing knowledge discovery in clinical domains, on the other hand, it poses some unique challenges for data analysis and interpretability, which call for new statistical techniques in both predictive and descriptive learning.

Machine learning (ML) algorithms, used in data science to process data that may exceed the capacity of the human brain, in order to make predictions or decisions without being explicitly programmed to do so (Obermeyer & Lee, 2017; Rajkomar *et al.*, 2019; Johnson *et al.*, 2016a), complement classical statistical tools. They are swiftly infiltrating many areas within the healthcare industry and biomedical domains for better informing the care of each patient. That is, decisions management, diagnoses and therapies are personalized on the basis of all known information about a patient, in real time, and incorporating lessons from a collective experience. An overview on successful biomedical applications which used ML, mostly *supervised* and *unsupervised* learning (Bishop, 2006) algorithms, is provided in Deo (2015) and Rajkomar *et al.* (2019).

As an alternative ML area, Reinforcement Learning (RL), offers a potential framework for tasks in which no initial data are provided and the algorithm has to learn by interacting with the surrounding environment in a sequential manner (Sutton & Barto, 2018; Bertsekas, 2019; Sugiyama, 2015). More specifically, in RL problems, at each time step of a sequential process, an *agent* interacts with its *environment*, performs *action(s)*, and, based on a *feedback* received from the environment for the selected action(s), learns, by *trial-and-error*, on how to take better actions in

order to maximize the cumulative feedback over time. Such distinctive feature offers a powerful solution in a variety of healthcare domains where the problem has a sequential nature (Chakraborty & Moodie, 2013; Yu *et al.*, 2019b; Gottesman *et al.*, 2019).

One of the emerging lines in both applied and methodological research within the domain of personalized medicine is the development of evidence-based (i.e., data-driven) adaptive interventions (AIs; Almirall *et al.*, 2014; Collins *et al.*, 2004); research line emerged from at least two different disciplines, i.e., RL and causal inference. The fundamental problem of AIs is to operationalize sequential decision making with the aim of optimizing individual outcomes by tailoring interventions to or by the individual patient over the course of a disease or program. This is the typical situation in clinical practice, in which a doctor needs to define a treatment regime depending on patients' characteristics, e.g., demographics, clinical conditions or previous response to a specific treatment regime (Murphy, 2003; Lavori & Dawson, 2004; Chakraborty & Murphy, 2014). Finding personalized therapies is a major challenge: it not only needs to handle "the right individual with the right treatement" but also "at the right time".

Reinforcement learning, perfectly resembling the AIs sequential problem, has been initially introduced into the clinical trial arena for discovering optimal *dynamic treatment regimes* (DTRs; Murphy, 2003; Lavori & Dawson, 2004; Chakraborty & Murphy, 2014) in life-threatening diseases such as cancer (Zhao *et al.*, 2009; Goldberg & Kosorok, 2012), and then spread to broader healthcare and behavioural areas including promoting physical activity (Yom-Tov *et al.*, 2017; Avila-Garcia *et al.*, 2019) and weight loss (Forman *et al.*, 2019; Pfammatter *et al.*, 2019) or managing the substance use (Goldstein *et al.*, 2017; Naughton, 2017). The *SMART Weight Loss Management* study (ClinicalTrials.gov Identifier: NCT02997943), for instance, seeks to develop an effective DTR strategy, to manage treatment for obesity. We report some details of the study, including the design, in Section 3.1. A more recent example (compared to DTRs), is given by *just-in-time adaptive interventions* (JITAIs; Tewari & Murphy, 2017; Nahum-Shani *et al.*, 2018), a growing standard in the emerging *mobile health* (mHealth) field (Istepanian *et al.*, 2007; Kumar *et al.*, 2013; Rehg *et al.*, 2017). JITAIs are AIs that use continuously collected data through mobile technology (e.g., wearable devices or smartphones) to adapt intervention components in real time for supporting behavior changes. MHealth applications (apps) for improving physical activity by delivering messages to users, represent a typical example. To discuss their use and benefits, we illustrate in Chapter 5 and Chapter 6 a mHealth app, named *DIAMANTE* (Avila-Garcia *et al.*, 2019), we developed for promoting physical activity. While in Chapter 5 we report the results of the preliminary DIAMANTE app-based study we conducted on a population of university students, in Chapter 6 we discuss the challenges we faced for implementing and adapting the same app on a population of patients with depression and diabetes, which is an ongoing study (ClinicalTrials.gov Identifier: NCT03490253; Aguilera *et al.*, 2020). We include details on both trial design and RL-based strategy. For the latter, we formulated the RL problem as a *multi-armed bandit* (MAB; Sutton & Barto, 2018; Lattimore & Szepesvári, 2020; Auer *et al.*, 2002a), a classic RL example.

Multi-armed bandit problems, have been extensively studied within statistics,

engineering and psychology from a long time. They have been introduced in biostatistics by Thompson (1933), and extensively studied under the heading *sequential design of experiments* (Robbins, 1952; Berry & Fristedt, 1985a; Lai, 1987). While these models are nowadays widely studied with completely different applications in mind, like online advertisement (Chapelle & Li, 2011) or recommender systems (Li *et al.*, 2010), there has been a surge of interest in the use of bandit algorithms for clinical trials (Villar *et al.*, 2015a), particularly given the increased attention payed by regulatory agencies to to *adaptive clinical trials* (FDA, 2019; Pallmann *et al.*, 2018), which use interim collected data to dynamically adjust the trial design. They have been proposed as a means to increase the efficiency of traditional *randomized clinical trials* (RCTs), not only benefiting future patients, but also trial participants, advancing patient care, and reducing costs (Bhatt & Mehta, 2016). MAB models might be particularly appropriate for designing adaptive clinical trials since the trade-off between clinical research and clinical practice can be seen as the well-known trade-off between exploration and exploitation in the context of RL. Under this pretext, the application of novel RL and MAB models tailored to clinical trials is being increasingly studied and shows great benefits compared to standard approaches. We review some of the proposed techniques in the context of adaptive-dose finding and response-adaptive randomization.

However, broader use of adaptive clinical trials, or, more generally, *adaptive experiments*, requires a better understanding of the trade-offs MAB algorithms make between the scientific and practical goal. For example, in a clinical trial, to achieve the scientific goal of a randomized trial such as testing whether a treatment is more effective compared to a control, typically, patients are randomized to treatments with equal and fixed probability. To achieve the practical (and more ethical) goal of assigning the best treatment more often, an algorithm that dynamically modifies the randomization probability of future patients, by using the evidence of previous patients' responses, would be preferred. A major barrier to adopting MAB algorithms for experimental designs is lack of clarity on how statistical analyses and inference are impacted (Rafferty *et al.*, 2019). Theoretical work suggests that adaptive data collection like the one used in bandit algorithms can induce bias in the estimates of means (Bowden & Trippa, 2017; Deshpande *et al.*, 2018; Nie *et al.*, 2018; Shin *et al.*, 2019) and that confidence intervals constructed from these statistics may not have correct coverage (Hadad *et al.*, 2019; Zhang *et al.*, 2020b). Both practical decisions and scientific research rests on knowing and controlling how frequently type-I error occurs, and having high power guarantees. In adaptively collected data, these measures represent a major challenge, and their poor understanding, and absence of robust inference and estimation in adaptively collected data, constitutes one of the main drivers that prevents the practical use of bandit strategies in clinical trials (Pallmann *et al.*, 2018; Burnett *et al.*, 2020).

Motivated by the existing challenges and open problems in many clinical and healthcare areas, where the use of RL and MABs have been argued to provide great benefits, we explore alternative ways of integrating statistical knowledge into analysis and improvement of machine learning techniques, with the aim to inspire the development and modification of theoretical frameworks and algorithms to better tackle these issues. We believe that there is scope for important practical advances in the use of RL and MABs in healthcare and behavioural areas, and with this work

we aim to make it easier for theoretical disciplines (RL and statistics) to join forces to assist clinical practice and medical discoveries and to develop the next generation of methods for AIs and ADs in healthcare.

1.1 Outline, Content and Contributions

A summary of the key contributions of this work in relation to its structure is given.

- Chapter 2. We begin by providing the biostatistical, and more generally the research community, with a mathematical formalization of the RL framework. We translate the key terminologies and approaches from an ML to a statistical perspective, assimilating also the different existing terminologies and notations into a coherent body of work. This offers a foundation to more easily conduct research in both theoretical and applied sciences.

- Chapter 3. Through a methodological review, we provide the general *panorama* of the applications of interest, i.e., developing adaptive interventions - in both clinical settings (dynamic treatment regimes) and behavioural sciences (mobile health applications) - and designing adaptive clinical trials. we show their natural formalization through RL, and illustrate the existing RL-based methods in their respective healthcare domain, discussing main similarities and differences among them in terms of their terminology of reference, trial design and main RL class.

- Chapter 4. Focusing on the specific application of adaptive clinical trials, particularly response-adaptive randomization, and motivated by the existing challenges in drawing inference from data adaptively collected by RL-based strategies, we quantify the extent of the problem and propose alternative ways of integrating statistical knowledge into improvement of this adaptive techniques. First, two ways of modifying the test statistic, using knowledge of the adaptive data collection nature, are explored; second, a modification of the algorithmic framework, for a more sensitive problem to both statistical inference as well as reward maximization, is considered.

- Chapter 5. Focusing on the specific application of JITAIs in the emerging mHealth area, we illustrate and discuss the benefits of a micro-randomized trial for promoting physical activity in university students through a text-messaging app, named DIAMANTE. We focus on the design of both the experimental trial and the adaptive RL-based algorithm, as well as the final results.

- Chapter 6. Using as motivating example the main (clinical) DIAMANTE study, we discuss the potential challenges that may arise when developing and designing JITAIs in mHealth, and highlight the crucial role of the biostatistical community in helping to overcome the existing issues.

Chapter 2

Reinforcement Learning Framework

Reinforcement learning is an area of machine learning concerned with determining optimal action selection policies in sequential decision making processes (Sutton & Barto, 2018; Bertsekas, 2019). The general framework is based on sequential interactions between a decision maker or *learning agent* and the *environment* it wants to learn about. More specifically, the agent and environment interact at each stage or time step $t \in \mathbb{N}$ of a sequence (here, we assume a discrete time space, even though it can be extended to the continuous-time case; Bertsekas & Tsitsiklis, 1996; Doya, 2000), in which the agent receives some representation of the environment's *state* or *context*, $X_t \in \mathcal{X}_t$, and on that basis makes a decision by selecting an *action* A_t from a set of admissible actions $\mathcal{A}_t$. As a result, one time step later, the environment responds to the agent's action by making a transition to a new state $X_{t+1} \in \mathcal{X}_{t+1}$ and (typically) providing a *reward* $Y_{t+1} \in \mathcal{Y}_{t+1} \subset \mathbb{R}$.

By repeating this process for each $t \in \mathbb{N} = \{0, 1, \dots, \}$, the result is a *trajectory* $\mathcal{T}$ of states visited, actions pursued, and rewards received:

$$\mathcal{T} \doteq \{(X_t, A_t, Y_{t+1})\}_{t \in \mathbb{N}} \,. \tag{2.1}$$

In a medical context, this trajectory can be viewed as the collection of information (e.g., covariates, treatments and responses to treatments) of a single patient i over the course of a disease. Note that in some settings there may be only one terminal reward (or 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 stages are taken to be 0. In other settings (e.g., multi-armed bandits, as we will see later in Section 2.0.1), the state is not considered, leading thus to a trajectory of actions and rewards only.

Define now $\mathbf{X_t} \doteq (X_0, \dots, X_t)$, $\mathbf{A_t} \doteq (A_0, \dots, A_t)$ and $\mathbf{Y_t} \doteq (Y_1, \dots, Y_t)$, and similarly $\mathbf{x_t}$, $\mathbf{a_t}$ and $\mathbf{y_t}$, where the upper and lower case letters denote random variables and their particular realization, respectively. Also define the *history* $\mathbf{H_t}$ (or *filtration* $\mathcal{F}_\mathbf{t}$) as all the information available at time t prior to agent's decision A_t, i.e $\mathbf{H_t} \doteq (\mathbf{X_t}, \mathbf{A_{t-1}}, \mathbf{Y_t})$; similarly $\mathbf{h_t}$. Stage t history's space, denoted by $\boldsymbol{\mathcal{H}_t}$, is therefore the product of $\mathbf{H_t}$ elements' spaces, i.e. $\boldsymbol{\mathcal{H}_t} = \mathcal{X}_0 \times \prod_{\tau=1}^{t} \mathcal{X}_\tau \times \mathcal{A}_{\tau-1} \times \mathcal{Y}_\tau$. Note that, by definition, $\mathbf{H_0} = X_0$.

We assume that these longitudinal histories (or equivalently the trajectories in (2.1) plus the final state) are sampled independently according to a distribution $P_{\boldsymbol{\pi}}^{\text{Full-RL}}$, with superscript clarified later in Section 2.0.1, given by:

$$P_{\boldsymbol{\pi}}^{\text{Full-RL}} \doteq p_0(x_0) \prod_{t \geq 0} \pi_t(a_t|\mathbf{h_t}) p_{t+1}(x_{t+1}, y_{t+1}|\mathbf{h_t}, a_t), \tag{2.2}$$

where:

- p_0 is the initial probability distribution specifying the initial state X_0.
- $\boldsymbol{\pi} \doteq \{\pi_t\}_{t \geq 0}$ represents the *exploration policy* and it determines the sequence of actions generated throughout the decision making process. More specifically, π_t maps histories of length t, $\mathbf{h_t}$, to a probability distribution over the action space $\mathcal{A}_t$, i.e. $\pi_t(\cdot|\mathbf{h_t})$. 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 $\mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}$. Sometimes, the action A_t to take at each time step t is uniquely determined by the history, therefore the policy is simply a function of the form $\pi_t \colon \boldsymbol{\mathcal{H}_t} \to \mathcal{A}_t$, or equivalently $\pi_t(\mathbf{h_t}) = a_t$. 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 characterize the dynamics of the environment. At each time $t \in \mathbb{N}$, the transition probability p_t assigns to each state-action-reward sequence $(\mathbf{x_{t-1}}, \mathbf{a_{t-1}}, \mathbf{y_t}) = (\mathbf{h_{t-1}}, a_{t-1})$ of the trajectory up to time $t-1$ a probability measure over $\mathcal{X}_t \times \mathcal{Y}_t$, i.e. $p_t(\cdot, \cdot|\mathbf{h_{t-1}}, a_{t-1})$. At each time t, the transition probability gives rise to:
  - $p_t(x_t|\mathbf{h_{t-1}}, a_{t-1})$, the *state-transition probability distribution* which represents the probability of moving to state x_t provided that a certain history $\mathbf{h_{t-1}}$ was observed up to time $t-1$ and that an action a_{t-1} was chosen in state x_{t-1}.
  - $Y_t = Y_t(\mathbf{H_{t-1}}, A_{t-1}, X_t)$, the *immediate reward function*. Generally, in RL, the immediate reward Y_{t+1} is conceptualized as a known function (rather than distribution) of the history $\mathbf{H_t}$, the current selected action A_t and the new state X_{t+1}; we thus, adapt our notation to $Y_{t+1} = Y_{t+1}(\mathbf{H_t}, A_t, X_{t+1})$. To give a concrete example, one can think of a dose-finding trial, where the level of toxicity is one of the covariates (or state variables), among the 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 potentially defined as a binary variable assuming value -1 if a toxicity level (X_{t+1}) higher than a certain value α is observed, and 0 otherwise.

The cumulative sum of immediate (future) rewards, or, more generally, a *discounted* version of it, is called *return* or *discounted return*. At time t, the *discounted return*, denoted by $\mathbf{R_t}$, an agent is going to receive over the future is defined as:

$$\mathbf{R_t} \doteq Y_{t+1} + \gamma Y_{t+2} + \gamma^2 Y_{t+3} + \cdots = \sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1}, \quad t \in \mathbb{N}. \tag{2.3}$$

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 infinte sum in (2.3) 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 maximizing immediate rewards, i.e. $\mathbf{R_t} = Y_{t+1}$. If $\gamma = 1$, the return is called *undiscounted* and it is well defined (finite) as long as the trajectory in (2.1) is finite, i.e. $t \in [0, T]$, with $T < \infty$, so that $\mathbf{R_t}$ is a sum of a finite number of elements (Sutton & Barto, 2018). When agent-environment interactions have a terminal stage $T < \infty$, the trajectory is also called *episode* and the agent has to face an *episodic task*. An episodic task is also known in the general RL framework as *finite-horizon task* if T is fixed and known in advance (e.g., in clinical trials), or *indefinite-horizon task* if T is not pre-specified and can be arbitrarily big (the typical case of EHRs). On the contrary, for $T = \infty$, the task is called *continuing task* or *infinite-horizon task* (Sutton & Barto, 2018).

Solving a reinforcement learning task means, roughly, learning an optimal way of choosing the set of actions or learning an *optimal policy*, so as to maximize the expected future return. However, in many sequential decision problems, the *target policy* we want to learn about might be different from the *exploration policy* $\boldsymbol{\pi}$ that generated the data. This happens for instance when we use trajectory samples generated from a policy which does not correspond to the policy of interest we want to estimate, for instance another trial. We call this target policy of interest *estimation policy* and denote it with $\mathbf{d}$. Thus, being concerned of this potential policy change, the RL problem at time t is to find an optimal policy $\mathbf{d_t^*} \doteq \{d_t^*\}_{\tau \geq t}$ such that

$$\boldsymbol{d_t^*} = \underset{\boldsymbol{d_t} \in \boldsymbol{\mathcal{D}_t}}{\arg\max}\, \mathbb{E}_{\boldsymbol{d}}[\mathbf{R_t}] = \underset{\boldsymbol{d_t}}{\arg\max}\, \mathbb{E}_{\boldsymbol{d}} \left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1} \right], \quad \forall t \in \mathbb{N}, \tag{2.4}$$

where the expectation is meant with respect to a trajectory distribution analogous to (2.2), say $P_{\mathbf{d}}$, where the fixed exploration policy $\boldsymbol{\pi}$ that generated the data is replaced by an arbitrary policy $\mathbf{d} \in \mathcal{D}$ we use to train the data, i.e., $\mathbb{E}_{\mathbf{d}} = \mathbb{E}_{P_{\mathbf{d}}}$. For ease of notation we also use $\mathbb{E}_{\boldsymbol{d}}$ for $\mathbb{E}_{\boldsymbol{d_t}}$: the time index is already incorporated in the argument.

For policy learning, various methods have been developed so far. These methods can be classified into *model-based reinforcement learning* and *model-free reinforcement learning*, where the term "model" indicates a model of the unknown environment, i.e., the transition probability distributions $\{p_t\}_{t \geq 1}$. We refer to Sutton & Barto (2018) and Sugiyama (2015) for the reader interested in an extensive overview. However, traditionally, in a broad range of literature and applications, by optimal policy we mean the one with the greatest *value*, i.e., the greatest expected return by following it when starting from a given state (*state-value* or simply *value*) or a given state-action pair (*action-value* or *Q-value*). Thus, efficiently estimating the value function is one of the most important component of almost all RL algorithms, and it occupies a central place in the medical decision making paradigm.

The stage t *state-value function* or *value function* of a fixed policy $\boldsymbol{d_t}$, maps a starting history $\mathbf{h_t}$ (with terminal state $X_t = x_t$) to the expected return. Formally,

$\forall t \in \mathbb{N}$ and $\forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}$, we denote it by $V_t \doteq V_{\boldsymbol{d_t}} : \boldsymbol{\mathcal{H}_t} \to \mathbb{R}$ and define it as

$$V_t(\mathbf{h_t}) \doteq V_{\boldsymbol{d_t}}(\mathbf{h_t}) \doteq \mathbb{E}_{\boldsymbol{d}}\left[\mathbf{R_t} | \mathbf{H_t} = \mathbf{h_t}\right] = \mathbb{E}_{\boldsymbol{d}}\left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1} \middle| \mathbf{H_t} = \mathbf{h_t}\right]. \tag{2.5}$$

To ensure that the conditional expectation in $V_t(\mathbf{h_t})$ is well defined, each history $\mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}$ should have positive probability ($\mathbb{P}(\mathbf{H_t} = \mathbf{h_t}) > 0$). Note that, by definition, at stage $t = 0$, $V_0(\mathbf{h_0}) = V_{\boldsymbol{d_0}}(x_0) \doteq V(x_0)$; while for the terminal stage, if any, the state-value function is 0.

Similarly, we define the stage t *action-value function* for policy $\boldsymbol{d_t}$, also known as *Q-value* or *Q-function*, as the expected return at time t, when starting from a history $\mathbf{h_t}$, taking an action a_t and following the policy $\boldsymbol{d_t}$ thereafter. Denoted it by $Q_t \doteq Q_{\boldsymbol{d_t}} : \boldsymbol{\mathcal{H}_t} \times \mathcal{A}_t \to \mathbb{R}$, we have that, $\forall t \in \mathbb{N}$, $\forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}$ and $\forall a_t \in \mathcal{A}_t$,

$$Q_t(\mathbf{h_t}, a_t) \doteq \mathbb{E}_{\boldsymbol{d}}\left[\mathbf{R_t} | \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right] = \mathbb{E}_{\boldsymbol{d}}\left[\sum_{\tau \geq t} \gamma^{\tau - t} Y_{\tau+1} \middle| \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right], \tag{2.6}$$

where, analogously to (2.5), $\mathbf{H_t}$ and A_t are randomly selected such that $\mathbb{P}(\mathbf{H_t} = \mathbf{h_t}) > 0$ and $\mathbb{P}(A_t = a_t) > 0$.

At stage t, the *optimal Q-function* $Q_t^* \doteq Q_{\boldsymbol{d_t^*}}$ and the *optimal value function* $V_t^* \doteq V_{\boldsymbol{d_t^*}}$ for policy $\boldsymbol{d_t}$ are defined as follows

$$Q_t^*(\mathbf{h_t}, a_t) \doteq \max_{\boldsymbol{d_t}} Q_t(\mathbf{h_t}, a_t), \quad \forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}, \forall a_t \in \mathcal{A}_t \tag{2.7}$$

$$V_t^*(\mathbf{h_t}) \doteq \max_{\boldsymbol{d_t}} V_t(\mathbf{h_t}) \doteq \max_{a_t \in \mathcal{A}_t} Q_t^*(\mathbf{h_t}, a_t), \quad \forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}. \tag{2.8}$$

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

$$d_t^*(\mathbf{h_t}) \in \arg\max_{\boldsymbol{d_t}} V_t(\mathbf{h_t}) = \arg\max_{a_t \in \mathcal{A}_t} Q_t^*(\mathbf{h_t}, a_t). \tag{2.9}$$

A fundamental property of value functions used throughout RL is that they satisfy particular recursive relationships. For any policy $\mathbf{d}$, the following consistency condition, known as Bellman equation for the value function, holds:

$$V_t(\mathbf{h_t}) = \mathbb{E}_{\boldsymbol{d}}\left[Y_{t+1} + \gamma V_{t+1}(\mathbf{h_{t+1}}) | \mathbf{H_t} = \mathbf{h_t}\right], \quad \forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}, \forall t \in \mathbb{N}. \tag{2.10}$$

It expresses the relationship between the value of a state and the values of its successor states: the value of the start state is equivalent to the value of the expected next state plus the expectation of the reward along the way. Based on this property and the definitions given in (2.7)-(2.8), for discrete state and action spaces the following important rules, known as Bellman optimality equations (Bellman, 1957), are satisfied:

$$V_t^*(\mathbf{h_t}) = \mathbb{E}\left[Y_{t+1} + \gamma V_{t+1}^*(\mathbf{H_{t+1}}) \mid \mathbf{H_t} = \mathbf{h_t}\right], \quad \forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}, \tag{2.11}$$

$$Q_t^*(\mathbf{h_t}, a_t) = \mathbb{E}\left[Y_{t+1} + \gamma \max_{a_t \in \mathcal{A}_t} Q_t^*(\mathbf{h_t}, a_t) \mid \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right], \quad \forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}, \forall a_t \in \mathcal{A}_t. \tag{2.12}$$

Here, the expectation 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 estimation of value functions recursively, from T backwards in time. In finite-horizon *dynamic programming* (DP) this technique is known as *backward induction*, and represents one of the main methods in for solving the Bellman equation.

2.0.1 Specific Formalizations of the RL Problem

The RL problem can be posed in a variety of different ways, depending on assumptions about the level of knowledge initially available to the agent. The framework is abstract and flexible and can be applied to many different (sequential) problems, by specifically characterizing 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 2, does not make any simplifying assumptions on the dependency between rewards, actions and states: by carrying over all the available history from the starting time, 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. To illustrate, consider a hospital admission scheduling problem (Kolesar, 1970), in which the decision (or the action) is represented by the number of daily admissions. In order to determine the optimal action, one may need to know the current (or at a certain time, e.g., daily) number of beds occupied, but neither the number of beds occupied at all the previous decision points, nor the set of all the previous actions. In other words, one may ignore the overall history and consider only the current state in the decision making process.

Alternatively, in some applied problems (e.g., indefinite-horizon problems), a full-RL formalization may be unfeasible and/or intractable for both estimation and inference purposes, requiring thus some forms of simplification in the distribution of the longitudinal histories. In JITAIs, for instance, the "just-in-time" nature of a decision making, assumes that the underlying decision rule is applied at the moment, without any severe computational time costs.

Common examples of specific formalizations of an RL problem, include *Markov decision processes* (MDPs) and *multi-armed bandit* (MAB) or contextual MAB problems. While here we discuss the MAB problem as a subclass of, or a specific way of formalizing, the RL problem (as in Sutton & Barto, 2018), we want to point out that, belonging to different research areas (RL is mostly associated to ML, while MABs to mathematics), some key researchers in the domain (see e.g., Lattimore & Szepesvári, 2020) distinguish between the two. One driver of this choice may be related to the major focus and attention to theoretical guarantees, e.g., optimal *regret* bounds, that MAB algorithms are expected to satisfy.

In what follows, we illustrate more in depth these two specific formalizations, starting with the MDPs, the main framework in indefinite-horizon DTRs problems. A graphical illustration of the different settings is preliminarly given in Fig. 2.1.

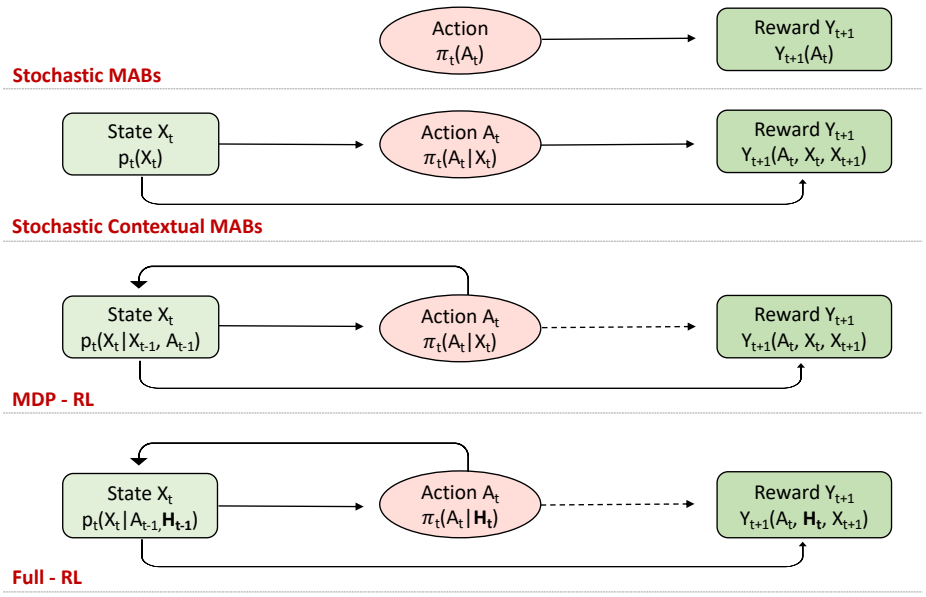

Figure 2.1. Main difference in terms of relationships between states, actions and rewards in a full RL (full-RL), MDP-based RL (MDP-RL), stochastic contextual MABs, and simple stateless or non-contextual MABs. In stochastic non-contextual MABs, immediate rewards depend on the current action only; in stochastic contextual MABs, immediate rewards depend on the current action and context; in MDP-RL, they depend on the current action and context as well as previous-time action and context; in full-RL, they depend on the entire history upon that time, including current action and state. The dashed line indicates a potential delayed effect in time of actions on the reward.

Markov Decision Processes

An MDP is a stochastic control process used to define environment's dynamics and to model the interaction between the agent and the controlled environment. It provides a mathematical framework for modeling decision making in situations where rewards are partly random and partly under the control of a decision maker (Puterman, 2014), and it is the most common setting assumed for an RL problem (Van Otterlo & Wiering, 2012).

What distinguishes an MDP-based RL (MDP-RL) from to the full-RL framework is the environment's random memory-less characteristic that informs the agent about its transition probabilities and guides the decision-making process. More specifically, assuming that the current state X_t contains all the information from the past history $\mathbf{H_{t-1}}$ (including also the current reward Y_t) that is meaningful to predict the future, it allows to ignore all the past histories when modelling next states and rewards. This property, known as *Markov property*, leads to a finite-size representation of the past, exemplifying all the trajectory distribution in (2.2) as follows:

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

and, exemplifying, thus, the entire optimization procedure required for computing

the optimal policy as reported in (2.4) or (2.9). Note that, under the Markov property, the agent's decisions can be entirely determined based on the current information only, as the latter fully determines environment's transition-probability distributions, i.e., $p_{t+1}(\cdot,\cdot|\mathbf{H_t}, A_t) = p_{t+1}(\cdot,\cdot|X_t, A_t)$, $\forall t \geq 0$. When transition probabilities $\{p_{t+1}\}_{t\geq 0}$ are also time independent, i.e., $p_{t+1} = p$, $\forall t \geq 0$ 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 previous stage information. As we will see later is Section 3.1.1, time-homogeneous MDPs were proposed in indefinite-time horizon DTRs, as they exemplify the problem by working with time-independent quantities which do not require a backward induction strategy.

While RL, including both Full-RL and MDP-RL, is typically formulated as a problem with states, actions, and rewards, with related transition rules, an exception is made for MAB problems, whose original formulation can be viewed as a *stateless* variant of RL.

Multi-Armed Bandits

MABs problems, often identified as a subclass of RL problems, have a long history in the statistical literature. They have been introduced in biostatistics by Thompson (1933) and, then, extensively studied under the heading *sequential design of experiments* (Robbins, 1952; Berry & Fristedt, 1985b; Lai, 1987).

Generally speaking, the MAB problem (also called the K- or N-armed bandit problem) is a problem in which a fixed limited set of resources must be allocated between competing choices in order to maximize the expected total reward over time. Each of the K choices provides a different reward, whose probability distribution is specific to that choice or action. 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 partial, for each time t the agent must trade-off between optimizing its decisions based on acquired knowledge up to time t (*exploitation*) and acquiring new knowledge about the expected payoffs of the other actions (*exploration*).

MABs strategies were originally proposed for solving stateless problems, in which the reward optimization task is based on actions only. Subsequently, a "stateful" variant of MABs, named *contextual* MAB (C-MAB), in which actions are associated with some signal, or *context*, was introduced. However, compared to Full-RL and MDP-RL, in contextual MABs, actions do not have any effect on next states. In addition, generally, there are no transition rules from one state to another in subsequent times, implying that states, actions and rewards can be treated as a set of separate events within time. The most typical assumption is that contexts $\{X_t\}_{t\in\mathbb{N}}$ are independent and identically distributed (i.i.d.) with some fixed, but unknown distribution. This means that action A_t at time t has an *in-the-moment* effect on the proximal reward Y_{t+1} at time $t+1$, but not on the distribution of future rewards $\{Y_\tau\}_{\tau\geq t+2}$, for which the i.i.d. property 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 problems is better known as *stochastic MABs*, in contrast with *adversarial MABs*, in which no independence restrictions on the sequence of rewards are made. In stochastic contextual MABs

the trajectory distribution is simplified as follows:

$$P_{\boldsymbol{\pi}}^{\text{C-MAB}} \doteq p_0(x_0) \prod_{t\geq 0} \pi_t(a_t|x_t) p_{t+1}(x_{t+1}, y_{t+1}|x_t, a_t)$$
$$= p_0(x_0) \prod_{t\geq 0} \pi_t(a_t|x_t) p_{t+1}(x_{t+1}) p_{t+1}(y_{t+1}|x_t, x_{t+1}, a_t),$$

with a further reduction in the non contextual MAB problem as follows:

$$P_{\boldsymbol{\pi}}^{\text{MAB}} \doteq \prod_{t\geq 0} \pi_t(a_t) p_{t+1}(y_{t+1}|a_t).$$

Note that, as 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, in which the time steps and states progression is not taken into account. MABs, thus, provides a simplified way (compared to both MDP-RL and full-RL) of formalizing the relationships between RL's components in time. A graphical summary of the different discussed RL frameworks is given in Fig. 2.1.

Similarly to the general RL, the MAB goal is to select the optimal arm A_t^* at each time t so that to maximize the expected return, alternatively expressed in the bandit literature in terms of minimizing the expected cumulative *regret*. Indeed, in (online) real-world problems, until we can estimate the optimal arms, we need to make repeated trials by pulling different arms. The loss that we incur during this learning phase (time/rounds spent for learning the best arm) represents what is called regret (i.e., how much we regret in not knowing/picking the optimal arm). More formally, at each round t, denoted with $A_t^* \doteq \arg\max_{a_t \in \mathcal{A}} \mathbb{E}(Y_{t+1}|X_t = \mathbf{x}_t, A_t = a_t)$ the optimal arm of that round, the expected regret is defined as the difference between the maximum possible expected reward and the expected reward resulting from the chosen action A_t, i.e.,

$$\text{reg}(t) \doteq \mathbb{E}(Y_{t+1}|X_t, A_t^*) - \mathbb{E}(Y_{t+1}|X_t, A_t), \tag{2.13}$$

Then, the goal of the learner is to minimize the cumulative sum of the regrets over a number T of steps, i.e., $\text{Reg}(T) \doteq \sum_{t=0}^{T} \text{reg}(t)$. In other words, the goal is to maximise the expected reward, not only after a total number of rounds, but also during the learning phase. More concretely, in a dose-finding problem as the one mentioned in Section 2.0.1, the aim is may be not only to minimize the sum of toxicities over time, but also to ensure that at each time t these have a proper upper bound, capable to limit extremely harmful adverse events. For this reason, as we will see later in Section 3.1.2, theoretical works on regret bounds occupy a central place in bandit literature.

Chapter 3

Review of RL Methods and Applications in Healthcare[1]

Abstract

Applications of reinforcement learning (RL) for supporting, managing and improving clinical and healthcare decisions are becoming increasingly popular. However, several challenges, which the biostatistical community may play a crucial role to help overcome, impact their use in real life. To bridge the statistics and RL communities, we aim to translate and extensively review the state-of-the-art research, with reference to both methodological and real-world studies. In this Chapter, we provide the first unified review on RL-based techniques used in medicine and other healthcare domains, including behavioural sciences, for supporting care and delivering optimized interventions.

We start by introducing and discussing the main applied areas which proposed RL as a potential solution for solving the related problem. With very few exceptions (e.g., *automated medical diagnosis*; Ling *et al.*, 2017), we classify them in two broad areas based on their goal: 1) to develop Adaptive Interventions (AIs), encompassing both dynamic treatment regimes (DTRs) and just-in-time adaptive interventions in mobile health (mHealth) - this is done in Section 3.1 - and, 2) to design adaptive clinical trials, including both exploratory and confirmatory phases - covered in Section 3.2. Then, we review the RL-methodologies which have been proposed within each area, providing a unified view (terminology and notation), and illustrate their potential benefits and drawbacks. Specifically for AIs, we show that, while a broad theoretical literature on constructing optimal DTRs with RL methods exists, their clinical application is still very limited. An opposite trend is registered in mHealth, where the number of applied real-world studies is continuously growing, but several methodological challenges (which will be discussed more in depth in Chapter 6) are poorly addressed. Within the domain of adaptive clinical trials designs, while applying RL (e.g., to adaptively adjust the randomization probabilities), can provide a huge benefit for the patients involved in the trial, it represents a major open problem in terms of inferential guarantees and results generalizability; this issue will

[1]Parts of the text of this chapter are extracted from the submitted/published manuscripts coauthored by the candidate and listed on page vii.

be illustrated in Chapter 4.

3.1 RL for Developing Adaptive Interventions

Adaptive Interventions (AIs) operationalize sequential decision making with the aim of optimizing individual outcomes and guiding practice over the course of a disease or, more generally, a program. They are represented by a sequence of decision rules that tailor the type, dose or delivery of intervention strategies, based on individuals' personal characteristics or progress, and repeatedly adjusted over time in response to ongoing performance (Almirall *et al.*, 2014; Nahum-Shani *et al.*, 2018). Existing frameworks for adaptive interventions (Collins *et al.*, 2004; Almirall *et al.*, 2014) highlight four components that play an important role in designing these interventions: (i) the decision points $t = 0, 1, \ldots$ specifying the time points at which a decision concerning intervention has to be made; and, at each point t, (ii) the intervention options belonging to the action space $\mathcal{A}_t$, (iii) the tailoring variables, i.e., $X_t \in \mathcal{X}_t$, and (iv) a decision rule d_t which links the tailoring variables to specific interventions. Intervention options correspond to different types, dosages (duration, frequency or amount; Voils *et al.*, 2012), or delivery options, as well as various tactical options (e.g., augment, switch, maintain), while tailoring variables capture information about an individual for a personalized decision making. An AI is a multistage process, wherein each stage corresponds to a period of time following a decision point in which the individual experiences an assigned intervention option. The assigned intervention option in at least one of the stages is tailored based on time-varying information about the participant, where "time-varying" refers to information that may change over time as a result of prior intervention stages (e.g., response to prior intervention, or motivational changes during the previous stage; Nahum-Shani *et al.*, 2017).

AIs are known by a variety of different names, with *adaptive treatment strategies* (Murphy, 2005a; Murphy *et al.*, 2007), *treatment policies* (Lunceford *et al.*, 2002; Dawson & Lavori, 2012; Wahed & Tsiatis, 2006), and *dynamic treatment regimes* (Murphy, 2003; Lavori & Dawson, 2004; Laber *et al.*, 2010; Chakraborty & Moodie, 2013; Laber *et al.*, 2014a), or *regimens*, being the most common ones. However, given the more generic nature and definition of AIs (sequential decision making formalization), we use this term to refer to a broad setting for selecting and personalizing interventions sequentially based on an individuals' time-varying characteristics, applicable, thus, not only in medical settings, but more generally in healthcare and other behavioural sciences, such as education (Nahum-Shani & Almirall, 2019). More specifically, we address two types of healthcare AIs in which RL methods have been employed: DTRs and JITAIs. We introduce them here, and discuss them more in depth in Sections 3.1.1 and 3.1.2, respectively.

With DTRs we refer to sequence of decision rules that dictate how to personalize treatments to patients, which typically has to be treated at multiple pre-defined stages, based on their evolving history (time-varying, dynamic state). In these settings, for each stage t, the actions or arms $A_t \in \mathcal{A}_t$ are treatments, the state $X_t \in \mathcal{X}_t$ is the set of patients' available information or covariates, and the reward $Y_t \in \mathcal{Y}_t$ is an intermediate outcome of interest. In some problems, there may only

be an end-of-study outcome of interest $Y = Y_{T+1}$ instead of multiple intermediate outcomes. For example, in the *attention deficit/hyperactivity disorder* (ADHD) Study (Pelham *et al.*, 2002) for evaluating the effects of a treatment on children with ADHD, the target outcome was school performance score at the end of study.

The set of decision rules $\mathbf{d} = \{d_t\}_{t\geq 0}$, or policies, is typically referred to as DTR, and each trajectory from the decision process corresponds to the complete history $\mathbf{H_t} \in \boldsymbol{\mathcal{H}_t}$ of baseline and time-varying covariates, assigned treatments, and observed outcomes of a single patient. TABLE 3.1 serves a table of equivalence for the different terminologies of reference in each setting, with a unified notation according to the general RL framework. Note that, while we report only the most common terminology adopted in each setting, the lexical borrowing is widely used across the different theoretical and applied domains. It is not rare to encounter for instance the term treatment policy instead of treatment regime in the DTR literature, or the term arm or treatment instead of intervention in JITAIs.

Table 3.1. Terminology of reference in reinforcement learning (RL), multi-armed bandits (MABs), dynamic treatment regimes (DTRs) and just-in-time adaptive interventions (JITAIs).

Notation	**Terminology**		
	RL/MABs	DTRs	JITAIs
i	Single Trajectory	Single Patient	Single User
t	Time/Round	Stage, Interval	Time
X	State/Context	Tailoring Variables*	Contextual Variables*
A	Action/Arm	Treatment	Intervention
Y	Reward	Intermediate, Distal Outcome	Proximal Outcome
$\mathbf{H}$	History/Filtration	History	Filtration
$\boldsymbol{\pi}, \mathbf{d}$	Policy	Treatment Regime	Policy
$\boldsymbol{\pi}^*, \mathbf{d}^*$	Optimal Policy	Optimal Treatment Regime	Optimal Policy

*Both tailoring and contextual variables represent the set of baseline and time-varying information that is used to personalize the decision making. Alternative terms such as covariates or features (that we use with a slightly different meaning as we discuss in Section 3.1.2) are also common.

An interesting aspect is the popular use of terms typical of specific RL methods in applications where these methods are used: see e.g., the similarity between JITAIs and MABs terms. We anticipate that methods for constructing JITAIs generally belong to the MAB framework, while in DTRs the prevalent class is full-RL, followed by MDP-RL proposed for indefinite-horizon DTRs problems. In fact, the underlying theory of DTRs, characterized by potential delayed and/or carry-over effects of a treatment over time, and importance of the evolving history of a patient for predicting future outcomes, requires an accurate consideration of previous stages information. Generally, the meaningful relationship between the different variables of a patient's history does not allow to simply or ignore the (state-)transition rules, making full-RL (and exceptionally MDP-RL) the ideal candidate. On the other hand, the behavioural theory of a momentary effect of the action on the proximal outcome underlying mHealth applications, makes MABs a more suitable framework compared

to full-RL and MDP-RL in this setting. In addition, the less computational burden of carrying through all the historical information, allows MABs strategies to be applied on a continuous time basis, e.g., every hour, and efficiently construct JITAIs.

With (mHealth) JITAIs we refer to sequence of decision rules which use continuously collected data through mobile technology (e.g., activity sensors, wearable devices, accelerometers or smartphones) to adapt intervention components in real time to support behavior change and promote health. The "just-in-time" support is based on considerations on whether and when the intervention is needed, and, in addition to the four key elements of a standard AI, a JITAI is characterized also by v) a distal outcome, i.e., the ultimate goal (typically a clinical outcome) the intervention is intended to achieve, and vi) the proximal outcomes, which define intermediate measures of the distal outcome, through which the intervention can be made (Tewari & Murphy, 2017; Nahum-Shani *et al.*, 2018). A typical example of JITAIs is a physical activity JITAI. Here, intervention options might be whether or not to send an activity encouraging message, the proximal outcome might be the number of steps the person walked in a temporal range after intervention was sent, and the context a set of user's variables such as GPS location, calendar busyness or heartrate. In JITAIs, we use the term interventions or intervention options for actions/arms, proximal outcome for the reward variable, and context for the set of tailoring variables. As we will see in Section 3.1.2, the problem in JITAIs is generally framed as a contextual MAB problem, more suitable in dynamic environments where context and options can change rapidly. This is the reason why the two frameworks of MABs and JITAIs share some terminology (see Table 3.1).

Data for building optimized JITAIs can be gathered through *randomized controlled trials* (RCTs; Collins *et al.*, 2005), *single-case experimental designs* (Dallery *et al.*, 2013; Dallery & Raiff, 2014), *factorial experiments* (Collins *et al.*, 2009), or, most notably, *micro-randomized trials* (MRTs; Klasnja *et al.*, 2015). In MRTs, individuals are randomized hundreds or thousands of times over the course of the study, and, in a typical multicomponent intervention study, the multiple components can be randomized concurrently, making micro-randomization a form of a sequential factorial design. The goal of these trials is to optimize mHealth interventions by assessing the causal effects of each randomized intervention component and evaluate whether the intervention effects vary with time or the individuals current context. To better understand the design and value of MRTs, the design of the *DIAMANTE* Study is presented in Figure 3.1. In this trial, based on the assigned study group (Static, Adaptive or Control), patients might be randomized every day to receive a combination of the different factors' levels, including different categories of motivational (Factor M) and feedback (Factor F) messages, and different time frames (Factor T). The adaptive optimization strategy of the *DIAMANTE* Study will be illustrated in Section 5.

For developing DTRs, two sources of data are commonly used: longitudinal observational studies and sequentially randomized trials, more specifically *sequential multiple assignment randomized trial* (SMART; Lavori & Dawson, 2000; Dawson & Lavori, 2012; Murphy, 2005a). While observational trials are the most common, SMARTs are experiencing a period of rapid growth and are currently the gold standard for developing DTRs (Lei *et al.*, 2012; Kasari *et al.*, 2014). A SMART design is characterized by multiple stages of treatment, each stage corresponding

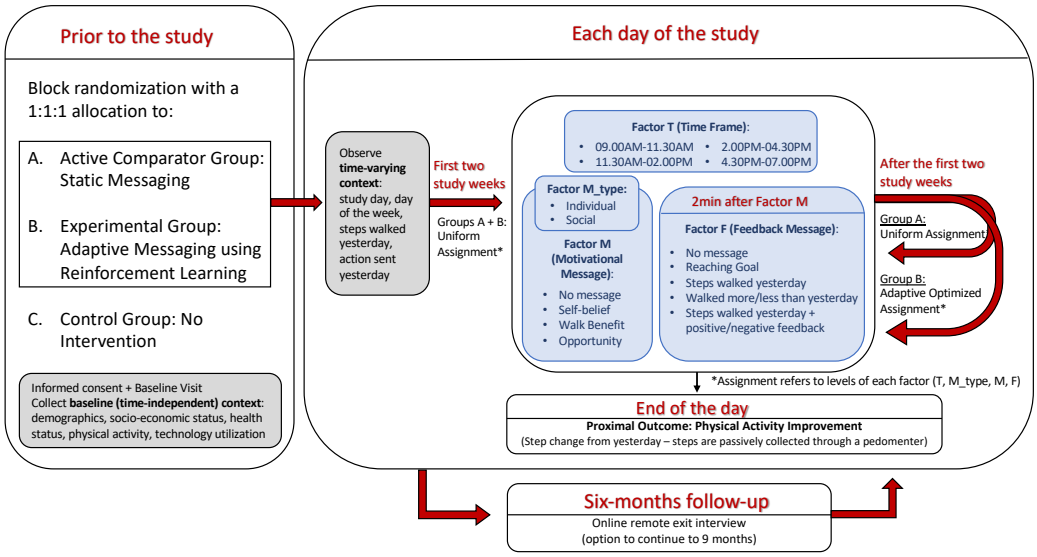

Figure 3.1. Schematic of the micro-randomized trial (MRT) design of the *DIAMANTE* Study (Aguilera *et al.*, 2020) for developing a smartphone messaging App to improve physical activity in patients with diabetes and depression.

to one of the critical decision time points. In this regard, it bears superficial similarity with *adaptive designs* (ADs; Bhatt & Mehta, 2016), in which certain trial features are allowed to change based on accumulating data. However, while in ADs each stage may involve different participants, and a treatment adaptation is made between-participants, in SMARTs the same participants move through multiple stages of treatment, involving a within-participant adaptation (Chakraborty & Murphy, 2014). In addition, in a SMART, unlike in an AD, typically, design elements such as the final sample size, randomization probabilities and treatment options are pre-specified. More aligned with traditional designs, the main goal of a SMART is to develop a good DTR that could benefit future patients, while ADs try to provide the most efficacious treatment to patients participating in the current trial. To give a concrete example, see Figure 3.2 for the schematic of the *Weight Loss Management* SMART design (Pfammatter *et al.*, 2019) mentioned in the Introduction. At program entry, all individuals are uniformly randomized to one of first-stage intervention options, either mobile app alone (App), for supporting self-monitoring of dietary intake and physical activity, or mobile app combined with weekly coaching (Coaching). Those achieving in 12 weeks < 0.5 lb weight loss on average per week, assessed by wireless scale at 2, 4, and 8 weeks, are classified as non-responders and re-randomized to one of two second-stage augmentation tactics: either modest augmentation, which consists of adding another mHealth component in the form of supportive text messages (TXT), or vigorous augmentation, consisting in adding supportive text messages combined with Coaching or meal replacement (MR). Responders continue the initial treatment option, and weight is assessed for all individuals in person at baseline, 3, 6, and 12 months, with weight change from baseline to 6 months being the primary outcome. Because different subsequent intervention options are considered for responders (continue) and non-responders (modest vs. vigorous augmentation), response status is embedded as a tailoring

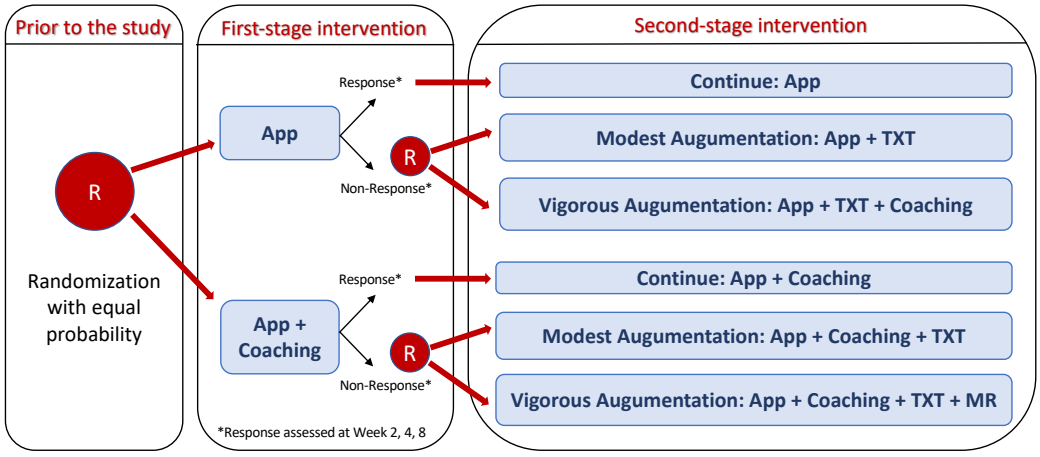

Figure 3.2. Schematic of the design of the sequential multiple assignment randomized trial (SMART) Weight Loss Management Study (Pfammatter *et al.*, 2019). App denotes a mobile app, TXT a supportive text message and MR meal replacement. Response is defined as a weight loss of at least 0.5 lb on average per week.

variable in this SMART by design. Such multistage restricted randomizations give rise to several DTRs that are embedded in the SMART.

SMARTs are considered the gold standard for developing DTRs, and compared to other types of randomized clinical trial designs, they offer tremendous advantages in terms of: (i) increased validity of analyses aimed at discovering when the effect of one intervention is enhanced by subsequent or prior interventions, (ii) increased validity of analyses aimed at discovering useful tailoring variables, and (iii) increased ability to reduce the impact of cohort effects (Lei *et al.*, 2012). Moreover, when it comes to its advantages compared to longitudinal observational studies, the discrepancy is even higher. Indeed, in observational studies the treatments are not randomized, and associations observed in the data (e.g., between treatment and outcome) may be partially due to the unobserved or unknown reasons why individuals receive differing treatments as opposed to the effects of the treatments. This adds an element of complexity to the problem of estimation and requires careful handling and additional assumptions for conducting causal inference. We review them in Section 3.1.1, where the *potential outcomes* framework, essential for estimating DTRs, is reported.

3.1.1 Dynamic Treatment Regimes

In medical research, DTRs define a sequence of treatments individually tailored to each patient based on baseline and time-varying (dynamic) state. In contrast with traditional single-stage treatments in which all individuals are assigned the same level and type of treatment, dynamic treatments explicitly incorporate the heterogeneity in treatment across individuals and the heterogeneity in treatment across time within an individual (Murphy, 2003), providing an attractive framework of personalized treatments in longitudinal settings. In addition, by treating only subjects who show a need for treatment, DTRs hold the promise of reducing non-compliance by subjects due to overtreatment or undertreatment (Lavori & Dawson, 2000; Collins *et al.*, 2001), and at the same time are attractive to public policy makers, allowing

a better allocation of public and private funds for more intensive treatment of the needy (Murphy, 2003).

Operationally, DTRs formalize the treatment decision making as a sequence of decision rules $\mathbf{d} = \{d_t\}_{t\geq 0}$, one per stage or time of intervention $t \in \mathbb{N}$. They dictate how to adapt the type and/or dosage, plus the timing of treatment, according to a patient's evolving conditions and treatments' history: each rule takes as input a patient's individual history $\mathbf{H_t} \in \boldsymbol{\mathcal{H}_t}$ up to stage t and outputs a recommended treatment $A_t \in \mathcal{A}_t$ from among the available feasible options $\mathcal{A}_t$. Throughout this section, we consider deterministic policies, which maps histories $\mathbf{h}$ directly into actions or decisions, i.e., $\mathbf{d}(\mathbf{h}) = \mathbf{a}$. The goal is to make these decisions so as to lead to the most beneficial expected outcome $Y_t \in \mathcal{Y}_t$ for an individual patient given history $\mathbf{H_t} \in \boldsymbol{\mathcal{H}_t}$. Assume that $\{Y_t\}_{t>0}$ are continuous variables coded so that higher values are preferred, the goal in then to find an *optimal DTR* $\mathbf{d}^* = \{d_t^*\}_{t\geq 0}$ that, if followed, yields the higher (most favourable), typically long-term mean outcome.

Methodology for constructing and evaluating optimal DTRs is of considerable interest within the domain of precision medicine, and comprises a growing body of research in both computer science and statistics (Chakraborty & Moodie, 2013; Laber *et al.*, 2014a). If from one side, DTRs problems, perfectly resembling the RL design, attracted the attention of ML researchers, from the other side, the necessity of quantifying causal relationships, rather than mere associations, called for the intervention of causal inference community. Indeed, the main challenge in DTR literature is that, since the underlying system dynamics are often unknown, inferring the consequences of executing a policy $\mathbf{d}$ and understanding the causal effects is not immediate.

Most of the current work in DTRs relies on the finite-horizon setting, and focus on the strongly related *offline learning* procedures, where one tries to identify the causal effect from finite observational data and causal assumptions about the data-generating mechanisms. Typically, in finite-horizon problems, estimation of the optimal DTR is obtain from these offline data, assuming we have access to the collection of observed trajectories for all patients. We recall from Section 2 that finite-horizon problems consider fixed length trajectories, with a terminal stage $T < \infty$ known in advanced. On the contrary, in indefinite or infinite-horizon problems, the number of stages is not *a-priori* specified and can be arbitrarily large or even infinite. This is particularly relevant for some chronic conditions, or those with very short time steps, in which patients have to be treated over the long term. However, despite its utility, only recently it has been addressed by DTR literature. Throughout this work I use the term "indefinite", and not "infinite", in line with the finite life expectancy of an individual.

Before delving into existing RL algorithms for estimating DTRs, we provide a taxonomy of the general RL methodologies. This will help the reader to better understand and move within this rich domain and will also serve as a guide for the development of the subsequent sections. Generally speaking, there are two fundamental learning mechanisms for deriving optimal policies in RL problems: *direct* and *indirect methods*. Direct methods seek optimal policies by directly looking for the optimal policy that maximises an objective (typically the expected return or value function) within a class of policies. On the contrary, indirect methods attempt, first, to estimate a value or Q-value function, and then to determine an optimal policy

based on the learned value function. This procedure is typically carried on through *dynamic programming* (DP) or *approximate dynamic programming* (ADP), by solving the Bellman equation of the value function and deriving the optimal action from it. Besides the above direct and indirect methods, there exists also the so called *actor critic* (AC) methods, which will be discussed more in depth in Section 3.1.2. AC algorithms keep separate, explicit representations of both value functions and policies and work on learning and improving both, providing, under certain assumptions, a unique architecture which unifies both direct and indirect methods (Guan *et al.*, 2019). In computer science literature, direct and indirect methods are sometimes referred to as *model-free* and *model-based* algorithms (Atkeson & Santamaria, 1997; Sutton & Barto, 2018). However, more subtle classifications (e.g., Guan *et al.*, 2019; Sugiyama, 2015) tend to make a clearer division between the two categories in the sense that direct/indirect are used for the learning process, while model-free/model-based refer to the data modelling assumptions. To illustrate, in Sugiyama (2015), model-free policy search is an equivalent of direct methods, while model-free policy iteration belongs to indirect strategies. A graphical understanding of the main classifications is provided in Figure 3.3. We adopt the general classification into

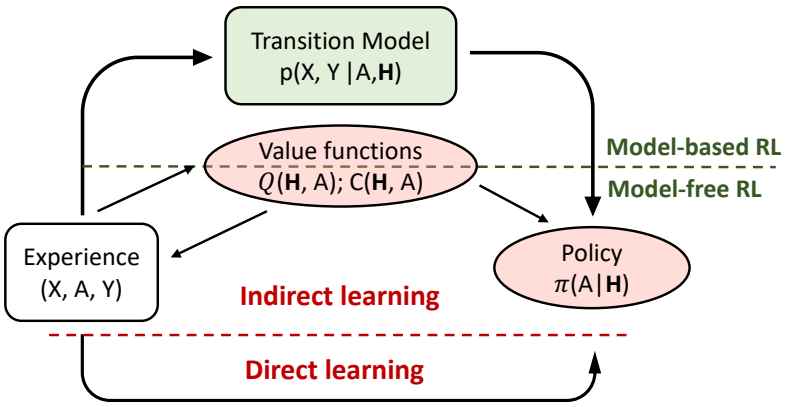

Figure 3.3. An illustration of different classes of RL methods and their learning process based on the classification in direct vs indirect learning and model-free vs model-based approach.

direct and indirect RL methods, in line with current DTR literature (Chakraborty & Murphy, 2014), and consider an additional subdivision into *fully-parametric*, *semi-parametric* and *non-parametric*, according to whether the data trajectory is fully-, partially- or not-parametrically modelled for estimation. Note that this sub-classification is analogous, but not identical, to the model-based and model-free classification, as the latter refers to a model for the entire environment (transition and reward distributions). As I will illustrate later, both indirect and direct methods can be parametric, but only indirect methods can be model-based.

In what follows, I review the existing techniques for estimating DTRs, starting with a brief digression on its origins within the causal inference literature and then focusing on the RL-based approach. Then, first the finite-horizon, which has been the focus of most of the DTR literature so far, is covered, and eventually the indefinite-horizon with some of the proposed *on-line learning* algorithms (e.g.,

Luckett *et al.*, 2020).

History of DTRs Estimation: from Statistics to Machine Learning

The study of DTRs, originated within causal inference, was pioneered by Robins (1986, 1994, 1997), with the introduction of *structural nested mean models* (SNMMs) and a number of estimating equation-based methods for finding optimal time-varying treatment regimes. SNMMs, which model the difference in the mean outcomes under different treatment regimes, rather than the full outcome model, were designed for estimating the joint effect of a sequence of treatments in the presence of a confounding variable (Robins, 1986). In this setting, standard regression methods, which attempt to estimate causal effects simultaneously are inappropriate, whether or not one adjusts for or conditions on the confounder. Over an extended period of time, the author introduced three basic approaches for finding optimal time-varying regimes in the presence of confounding variables: the parametric *G-formula* or *G-computation* (Robins, 1986), *structural nested models* (SNMs), which include SNMMs as a subclass, with the associated method of *G-estimation* (Robins, 1989, 1992, 1994), and *marginal structural models* (MSMs) with the associated method of *inverse probability of treatment weighting* (IPTW; Robins, 2000)). In spite of advantages and strong connections with popular estimation methods, SNMMs and G-estimation have not become as popular as MSMs and IPTW methods. A discussion on the possible reasons, with an accurate overview of the models and estimation methods as developed by Robins, can be found in Vansteelandt *et al.* (2014), who use the appellative "partially realized promise" referring to SNMs and G-estimation.

A number of methods have then been proposed within statistics, including frequentist and Bayesian likelihood-based approaches (Thall *et al.*, 2000, 2002, 2007), or methods based on multiple imputation for estimating and comparing all potential outcomes (Lavori & Dawson, 2000). However, all these methods first, infer the data-generation process via a series of parametric conditional models, then estimate the optimal DTRs based on the inferred data distributions. These approaches can easily suffer from model misspecification due to the inherent difficulty of modeling accumulative time-dependent and high-dimensional information in the models (Zhao *et al.*, 2015). The first semi-parametric method for estimating the optimal DTR was proposed by Murphy (2003), immediately followed by Robins (2004), who introduced two alternative approaches based on G-estimation and SNMMs, which generalize the approach of Murphy (2003) (Moodie *et al.*, 2007). These methods use approximate dynamic programming, where "approximate" refers to the use of an approximation of the value or Q-function, to estimate the optimal DTR. Thus, they can be considered as the first prototypes of RL-based approaches in the DTR literature.

Machine learning methods represent an alternative approach to estimating DTRs that have gained popularity due in part to their avoidance of having to completely model the underlying generative distribution. The main bridge connecting statistics and RL, previously confined to the computer science and control theory literature, was provided by the work of Murphy (2005b), who adapted Q-learning (Watkins, 1989; Sutton & Barto, 2018) to DTRs estimation, and derived an upper bound on its generalization error. It is based on ADP, and has been evaluated with parametric, semi-parametric, and non-parametric strategies (Murphy, 2005b; Chakraborty &

Moodie, 2013; Chakraborty & Murphy, 2014; Laber *et al.*, 2014b) for modelling the conditional mean outcome, i.e, the Q-function introduced in (2.6). Q-learning and the semi-parametric techniques of Murphy (2003) and Robins (2004) are indirect methods: optimal DTRs are obtained by indirectly modelling and estimating optimal value functions with ADP. On the contrary, IPTW-based techniques (Robins, 2000; Murphy *et al.*, 2001; Wang *et al.*, 2012), belong to direct methods, as they avoid the need for postulating a (conditional) outcome model (Zhao *et al.*, 2012).

An overview of the existing methods is provided in Figure 3.4. They are classified based on the main taxonomy adopted in this paper, i.e., direct vs indirect RL methods, and the nature of the assumed model for the data trajectory (fully-parametric, semi-parametric and non-parametric). Before going into these algorithms, it is essential

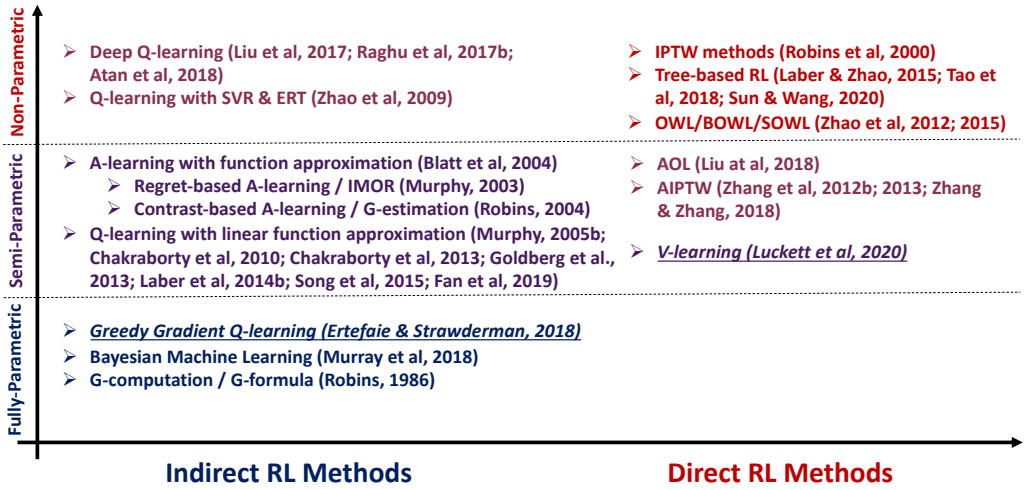

Figure 3.4. Schematic overview of RL-based techniques for estimating optimal DTRs. Methods are primarily classified according to the taxonomy we adopt in this manuscript, i.e., indirect (optimal policy is obtained by solving the Bellman equation, DP or ADP) vs direct (optimal policy is obtained by directly maximizing an objective function). Then an addition diversification is based on the parametric vs non-parametric nature of the likelihood: when the latter or parts of it are modelled by a parametric model, they are said to be parametric. Underlined methods refer to algorithms proposed in indefinite-horizon problems.

to mention that, in order to allow the quantification of treatment effects in DTRs and construct estimands of interest, Neyman (1923) and later Rubin (1974) laid the foundations of the so called *potential outcomes* or *counterfactual framework*, which, with the extension of (Robins, 1986) in both observational and randomized studies, represents the pillar of modern causal inference.

Potential Outcomes Framework

By far, the most popular approach to mathematically defining a causal effect is based on *potential outcomes*, or *counterfactuals*. With potential outcomes we refer to the set of all possible values of a status or outcome variable that would be achieved, if perhaps contrary to fact, the patient had been assigned to different treatments. In a simple one-stage RCT in which subjects can receive either treatments a and a', the

set of (unobserved) potential outcomes for an individual with baseline information X_0, is given by $\left(X_1^a, Y_1^a, X_1^{a'}, Y_1^{a'}\right)$, with $Y_1^a \doteq Y_1(X_0, a, X_1^a)$.

In order to define what we mean by a causal effect, for each individual (or subject, or unit) we assume the existence of the potential outcomes, $Y_1^a, Y_1^{a'}$, corresponding to what value the outcome would take if we did assign a or a', respectively. Then, to calculate the causal effect on a given individual we would need to somehow to compute the so called *individual-level causal parameter* given by $Y_1^a - Y_1^{a'}$. However, since we cannot observe all the potential outcomes on the same patient, typically *population-level causal parameter* (e.g., $\mathbb{E}[Y_1^a] - \mathbb{E}[Y_1^{a'}]$) are considered instead. In order to connect the potential outcomes with observed data, ensuring $\hat{\mathbb{E}}[Y_1|A = a]$ is an unbiased estimate of $\mathbb{E}[Y_1^a]$, the following assumptions about the assignment mechanism must hold.

1. *Stable unit treatment value assumption* (SUTVA), which assumes that each participant's potential outcome is not influenced by the treatment applied to other participants (Rubin, 1978, 1980). This assumption connects the potential outcomes to the observed data such that, for each t, $X_{t+1}^{\mathbf{a_t}} = X_{t+1}(\mathbf{a_t}) \doteq X_{t+1}$ and $Y_{t+1}^{\mathbf{a_t}} = Y_{t+1}(\mathbf{a_t}) \doteq Y_{t+1}$, when regime $\mathbf{a_t}$ is actually followed. This agreement between potential outcomes under the observed treatment and the observed data is known as *axiom of consistency*.

2. *No unmeasured confounders* (NUC), which states that conditional on the patient's history $\mathbf{H_t}$ up to time t, the treatment assignment A_t at time t is independent of future potential outcomes of the individual (Robins, 1997). That is, for any possible regime $\mathbf{a}$,

$$A_t \perp \left(X_{t+1}^{\mathbf{a_t}}, Y_{t+1}^{\mathbf{a_t}}, X_{t+2}^{\mathbf{a_{t+1}}}, Y_{t+1}^{\mathbf{a_{t+1}}}, \dots\right) | \mathbf{H_t}, \quad \forall t \in \mathbb{N}.$$

   This assumption always holds under either complete or sequential randomization, including SMART designs, but must be evaluated on subject matter grounds in observational studies.

3. *Positivity*, which defines the set of *feasible* regimes so that for every covariate-treatment history up to time t that has a positive probability of being observed, there must be a positive probability that the corresponding treatment dictated by the treatment regime will be observed (Robins, 1994). Formally, if we denote with $\boldsymbol{\pi}$ the probability distribution of actions given the history, a feasible regime $\mathbf{d}(\mathbf{h}) = \mathbf{a}$ satisfies

$$\pi_t(d_t(\mathbf{H_t})|\mathbf{H_t} = \mathbf{h_t}) > 0, \quad \forall \mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}, \forall t \in \mathbb{N}. \tag{3.1}$$

   That is, feasibility requires some subjects to follow regime $\mathbf{d}$ to guarantee non-parametric estimation of its performance.

   Please note that notation "π" is not arbitrary, it perfectly translates the notion of "exploration policy" introduced in Section 2 meant for the action process generation, and in a case of a randomized trial it consists of the randomization probabilities.

Under the consistency, sequential randomization and positivity assumptions, the conditional distributions of the observed data are the same as the conditional distributions of the potential outcomes. It follows that an optimal treatment regime may be obtained using the observed data.

RL Methods for Finite-Horizons Problems

Most of the existing methods in the DTRs literature, including those presented in Section 3.1.1, fall in the finite-horizon setting, as they are designed to optimize a utility function over a fixed period of time, say T. More specifically, given a finite-horizon trajectory $\mathcal{T} \doteq \{(X_0, A_0, Y_1, \ldots, X_T, A_T, Y_{T+1})\}$, with X_0 some pre-treatment information, $X_1, \ldots, X_T$ the evolving information, $A_0, \ldots, A_T$ the assigned treatments, and $Y_1, \ldots, Y_T$ the intermediate and the final (Y_{T+1}) outcomes, a sample (or *batch*) of N finite-horizon available patients' trajectories, each of the above form, are used for estimating an optimal DTR, which we denote with $\mathbf{d}^* = \{d_t^*\}_{t\geq 0}$. The problem conforms to what is known as *batch-mode* RL in computer science.

Indirect methods

With indirect methods we refer to a class of methods that focus on estimating an optimal objective function (typically, an expectation of the outcome variable such as the Q-function), and then get the associated policy, rather that directly looking for an optimal policy (see Figure 3.3). For learning and estimating the optimal value at each stage t, indirect RL methods are mainly based on iterative methods such as DP and ADP. These include Q-learning Murphy (2005b), where the conditional mean outcome is modelled, and other approaches, which we generally term *Advantage-learning (A-learning)*, which model contrasts of conditional mean outcomes. The latter has as examples the SNMMs with the G-estimation proposal of Robins (2004). and with the *iterative minimization of regrets* (IMOR) of Murphy (2003). We discuss them later in this section. To provide the reader with a unified overview, making at the same time a clear distinction between the two different procedures (even if based on an equivalent model formulation), we will consider the terms *Contrast-based A-learning* and *Regret-based A-learning*, to refer to Robins's and Murphy's works, respectively. This is done in line with the work of Schulte *et al.* (2014), which will be taken as a guide for discussing more in depth A-learning and compare it with the widely used technique of Q-learning.

Traditional statistical likelihood-based methods (Thall *et al.*, 2000, 2002), including the parametric G-computation (Robins, 1986) and Bayesian methods (Thall *et al.*, 2007), also fall into this category. We point to Tsiatis *et al.* (2019) and Vansteelandt *et al.* (2014) for readers interested in these traditional approaches.

Q-learning with function approximation. In Section 2 we introduced the main quantities of interest in RL problems, i.e., the value functions (see 2.5 and 2.6), and we showed that a powerful property of value functions used throughout RL, is that they satisfy particular recursive relationships between the value of a state and the values of its successor states, as reported in equation (2.10). This property is of fundamental importance, as it allows computation of optimal state-values $V_t^*(\mathbf{h_t})$ and Q-functions $Q_t^*(\mathbf{h_t}, a_t)$ at any time t, by solving the Bellman optimality equation

of DP (Bellman, 1957), also known as backward induction in finite-horizon problems. Specifically for the Q-function, given a model of the environment's dynamics, at any time t, for all $a_t \in \mathcal{A}_t$ and $\mathbf{h_t} \in \boldsymbol{\mathcal{H}_t}$, with discrete state and action spaces, we recall that an optimal Q-function $Q_t^*(\mathbf{h_t}, a_t)$ could be obtained as:

$$Q_t^*(\mathbf{h_t}, a_t) = \mathbb{E}\left[Y_{t+1} + \gamma \max_{a_t \in \mathcal{A}_t} Q_t^*(\mathbf{h_t}, a_t) \mid \mathbf{H_t} = \mathbf{h_t}, A_t = a_t\right].$$

However, the task is typically impossible to achieve due to unknown transition probability distributions (in DP we need an underlying model for the environment), memory and computations constraints related to the iterative procedure. RL algorithms that do not need an underlying model are known as *temporal-difference* (TD) learning and they constitute the core of modern RL, with *Q-learning* (Watkins, 1989) representing one of the most popular (off-policy) TD approaches. One fundamental component of TD-learning is the incremental implementation, which requires less memory for estimates and less computation. The general idea is that, for each time step $t \in [0, T]$, a new estimate is obtained based in part on an old previously learned estimate:

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

The constant α_t determines to what extent the newly acquired information will override the old information, that is, how fast learning takes place: a factor of 0 will make the learner not learn anything, while a factor of 1 would make the learner fully update based on the most recent information. The discount factor γ, introduced in Section 2, balances a learner's immediate rewards and future rewards, and in a finite horizon problem is generally set to one.

Under some appropriate and rigorous assumptions, Q_t has been shown to converge to the optimal Q-function Q_t^* with probability 1 (Watkins, 1989; Jaakkola *et al.*, 1994; Tsitsiklis, 1994). However, this simple approach is practical in a small number of problems because it can require many thousands of training iterations to converge in even modest-sized problems. In addition, it represents value functions in arrays, or tables, based on each state and action. Thus, large state spaces will lead not just to memory issues for large tables, but also to time problems needed to fill them accurately. A powerful, scalable way of generalizing this *tabular Q-learning* and to overcome the computational burden, involves *function approximation* (FA). We call this approach *Q-learning with function approximation*. The main idea of Q-learning with FA is first, to estimate the Q-function using an approximator, e.g., regression models, neural networks or decision-trees, and then to derive the estimated policy based on the estimated Q-function. More specifically, we start by assuming an approximation space for each of the t-th Q-functions, e.g., $\mathcal{Q}_t \doteq \{Q_t(\mathbf{h_t}, a_t; \theta_t) : \theta_t \in \Theta_t\}$, with parameter space Θ_t being a subset of the Euclidean space. According to the results shown in Section 2, estimating an optimal stage t policy is equivalent to estimate an optimal Q-function, or in this case, an optimal parameter $\hat{\theta}_t$, i.e.,

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

Then, according to Bellman optimality, we estimate an optimal regime $\hat{\mathbf{d}}^* = (d_0^*(x_0;\hat{\theta}_0), d_1^*(\mathbf{h_1};\hat{\theta}_1), \dots, d_T^*(\mathbf{h_T};\hat{\theta}_T))$ by recursively estimating Q_t^* backwards through time $t = T, T-1, \dots, 0$ (Bather, 2000). Formally, defined $Q_{T+1}^* \doteq 0$, we proceed as follows:

$$\begin{aligned} Q_T^*(\mathbf{h_T}, a_T;\hat{\theta}_T) &\doteq \hat{\mathbb{E}}[Y_{T+1}|\mathbf{H_T} = \mathbf{h_T}, A_T = a_T] \\ d_T^*(\mathbf{h_T};\hat{\theta}_T)) &= \arg\max_{a_T \in \mathcal{A}_T} Q_T^*(\mathbf{h_T}, a_T;\hat{\theta}_T) \\ Q_{T-1}^*(\mathbf{h_{T-1}}, a_{T-1};\hat{\theta}_{T-1}) &\doteq \hat{\mathbb{E}}[Y_T + \max_{a_T \in \mathcal{A}_T} Q_T^*(\mathbf{h_T}, a_T;\hat{\theta}_T)|\mathbf{H_{T-1}} = \mathbf{h_{T-1}}, A_{T-1} = a_{T-1}] \\ d_{t-1}^*(\mathbf{h_{t-1}};\hat{\theta}_{t-1})) &= \arg\max_{a_{T-1} \in \mathcal{A}_{T-1}} Q_{T-1}^*(\mathbf{h_{T-1}}, a_{T-1};\hat{\theta}_{T-1}) \\ &\dots \\ Q_0^*(x_0, a_0;\hat{\theta}_0) &\doteq \hat{\mathbb{E}}[Y_1 + \max_{a_1 \in \mathcal{A}_1} Q_1^*(\mathbf{h_1}, a_1;\hat{\theta}_1)|X_0 = x_0, A_0 = a_0] \\ d_0^*(x_0;\hat{\theta}_0)) &= \arg\max_{a_0 \in \mathcal{A}_0} Q_0^*(x_0, a_0;\hat{\theta}_0). \end{aligned} \tag{3.2}$$

We sometimes refer to this procedure as *batch Q-learning*, as learning occurs only after the collection of a set of N trajectories. The procedure, with a generic FA, is illustrated in Algorithm 1.

Algorithm 1: Q-learning with Function Approximation (Murphy, 2005b)

Input: Time horizon T, action and state spaces $\mathcal{A}$, $\mathcal{X}$, approximation space for the Q-functions $\mathcal{Q}_t \doteq \{Q_t(\mathbf{h_t}, a_t;\theta_t) : \theta_t \in \Theta_t\}$, for all $t = 0, \dots, T$.

Initialization: Stage $T+1$ optimal Q-function, for convenience it is typically set to $Q_{T+1}^*(\mathbf{h_{T+1}}, a_{T+1};\hat{\theta}_{T+1}) = \hat{\mathbb{E}}[Y_{T+1}|\mathbf{H_T} = \mathbf{h_T}, A_T = a_T] = 0$.

for $t = 0, 1, 2, \dots T$ **do**

Q-function parameters' estimate: get updated estimates $\hat{\theta}_{T-t}$ backwards by minimizing a loss, e.g., MSE, ($\mathbb{P}_N$ is the empirical mean over N trajectories)

$$\begin{aligned} \hat{\theta}_{T-t} \in \arg\min_{\theta_{T-t} \in \Theta_{T-t}} \mathbb{P}_N[Y_{T-t+1} + \max_{a_{T-t+1} \in \mathcal{A}_{T-t+1}} Q_{T-t+1}^*(\mathbf{h_{T-t+1}}, a_{T-t+1};\hat{\theta}_{T-t+1}) \\ - Q_{T-t}^*(\mathbf{h_{T-t}}, a_{T-t};\theta_{T-t})]^2. \end{aligned} \tag{3.3}$$

Optimal policy estimate: get the $(T-t)$-time optimal regime estimate as the one that maximises the optimal $(T-t)$-time Q-function estimate

$$d_{T-t}^*(\mathbf{h_{T-t}};\hat{\theta}_{T-t}) = \arg\max_{a_{T-t} \in \mathcal{A}_{T-t}} Q_{T-t}^*(\mathbf{h_{T-t}}, a_{T-t};\hat{\theta}_{T-t}) \tag{3.4}$$

end for

Several Q-learning function approximators have been proposed in literature, including linear regression, decision-trees or neural networks. As Q-functions are

conditional expectations, the first natural approach to model them is through regression models. Letting $\theta_t \doteq (\beta_t, \psi_t)$, Chakraborty & Moodie (2013) proposed stage specific optimal Q-functions to be parametrized as

$$Q_t^*(\mathbf{H_t}, A_t; \beta_t, \psi_t) = \beta_t^T \mathbf{H_{t0}} + (\psi_t^T \mathbf{H_{t1}}) A_t, \quad t \in [0, T], \tag{3.5}$$

where $\mathbf{H_{t0}}$ and $\mathbf{H_{t1}}$ are two (possibly different) vector summaries of the history $\mathbf{H_t}$, with $\mathbf{H_{t0}}$ denoting the "main effect of history" and $\mathbf{H_{t1}}$ denoting the "treatment effect of history". The collections of variables $\mathbf{H_{t0}}$ are often termed *predictive*, while $\mathbf{H_{t1}}$ are said *prescriptive* or *tailoring variables*. Parameters $\hat{\theta}_t \doteq (\hat{\beta}_t, \hat{\psi}_t)$ are obtained by solving suitable estimating equations such as *ordinary least squares* (OLS) or *weighted least squares* (WLS). Given a sample $\left\{X_{0i}, A_{0i}, Y_{1i}, \ldots, X_{Ti}, A_{Ti}, Y_{(T+1)i}, X_{(T+1)i}\right\}_{i=1}^N$ of i.i.d. trajectories, WLS (whose choice might be dictated by heteroschedastic errors), will estimate $\hat{\theta}_t$ by solving

$$\begin{aligned} 0 = \sum_{i=1}^N & \frac{\partial Q_t^*(\mathbf{H_{ti}}, A_{ti}; \theta_t)}{\partial \theta_t} \Sigma_t^{-1}(\mathbf{H_{ti}}, A_{ti}) \\ & \times [Y_{(t+1)i} + \max_{a_{(t+1)i} \in \mathcal{A}_{(t+1)i}} Q_{t+1}^*(\mathbf{H_{(t+1)i}}, a_{(t+1)i}; \hat{\theta}_{t+1}) - Q_t^*(\mathbf{H_{ti}}, A_{ti}; \theta_t)], \end{aligned}$$

where Σ_t is a working variance model. Taking Σ_t to be a constant yields the OLS estimator. Now, substituting $\{\hat{\theta}_t\}_{t\in[0,T]} \doteq \{\hat{\beta}_t, \hat{\psi}_t\}_{t\in[0,T]}$ in the process (3.2) yields an estimator for the optimal treatment regime as follows

$$\hat{\mathbf{d}}^* = (d_0^*(x_0; \hat{\theta}_0), d_1^*(\mathbf{h_1}; \hat{\theta}_1), \ldots, d_T^*(\mathbf{h_T}; \hat{\theta}_T)).$$

As noticed first by Robins (2004) for G-estimation, and then by Chakraborty *et al.* (2010) for Q-learning, the treatment effect parameters at any stage prior to the last, can be non-regular under certain longitudinal distributions of the data. Q-learning, for instance, involves modeling non-smooth, non-monotone functions of the data, which complicates both regression function and inference. Particularly, with (3.5) as model for the Q-functions, $\hat{\psi}_t$ is a *non-regular* estimator, and inferential problems arise when $\hat{\psi}_t^T \mathbf{H_{t1}}$ is close to zero, as non-differentiable in that point. To solve this issue, Chakraborty *et al.* (2010), adapting previous work in the context of G-estimation (Moodie & Richardson, 2010), proposed two alternative ways of shrinking or thresholding values of $\hat{\psi}_t^T \mathbf{H_{t1}}$ near zero. In a similar spirit, Song *et al.* (2015); Goldberg *et al.* (2013) proposed minimizing a penalized version of the objective in the first step of Q-learning, where the penalty is given by a function $p_\lambda(|\psi_t^T \mathbf{H_{t1}}|)$ with tuning parameter λ, while Fan *et al.* (2019) introduced the *smoothed Q-learning* dictated by the use of a modified version of $\hat{\psi}_t^T \mathbf{H_{t1}}$ in (3.5), given by $(\hat{\psi}_t^T \mathbf{H_{t1}}) K_\alpha(\hat{\psi}_t^T \mathbf{H_{t1}})$. Here, $K_\alpha(x) \doteq K(x/\alpha)$, with $\alpha > 0$ a smoothing parameter and $K(\cdot)$ a kernel function that admits a probability density function. Another proposal for conducting inference for the estimated Q-function parameters arised in Chakraborty *et al.* (2013), where a general method for bootstrapping under non-regularity, i.e., *m-out-of-n bootstrap* was presented. Subsequently, Laber *et al.* (2014b) derived a new *interactive Q-learning* method, where the maximization step is delayed, by adding an additional step between (3.3) and (3.4). This enables all modeling to be performed before the non-smooth, non-monotone transformation.

In addition to the challenges in development of statistical inference, it is important to recognize that the estimated regime $\hat{\mathbf{d}}^*$ may not be a consistent estimator for the true optimal regime $\mathbf{d}^*$, unless all the models for the Q-functions are correctly specified (Schulte *et al.*, 2014). The model in (3.5), for example, is quite simple due to its linearity, thus prone to misspecification. In order to address this problem, several FA alternatives have been proposed. Zhao *et al.* (2009), for instance, discussed and evaluated through simulations the potential of Q-learning with functions approximated by *support vector regression* (SVR) and *extremely randomized trees* (ERTs), techiques adopted from the ML literature. In SVR, the input data $\mathbf{x}_i \doteq \{x_{it}, a_{it}\}_{t=0,\dots,T}$, for $i = 1, \dots, N$, are mapped into a feature space by a non-linear transformation Φ, which guarantees that any data set becomes arbitrarily separable as the data dimension grows (Vapnik *et al.*, 1997). Then a *hyperplane* $f(\mathbf{x}_i)$, equivalent to the Q-function, is fitted to the mapped data, i.e., $f(\mathbf{x}_i, \mathbf{a}_i) = \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{a}_i) + b$. The goal of SVR is to find a function that deviates from the labels $\{y_{it}\}_{t=0,\dots,T}$ by a value no greater than ϵ for each training point; it thus solves an optimization problem of the form:

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi'} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i'), \\ \text{subject to } (\mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{a}_i) + b) - y_i \leq \epsilon + \xi_i, \\ y_i - (\mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{a}_i) + b) \leq \epsilon + \xi_i', \\ \xi_i, \xi_i' \geq 0, \quad i = 1, \dots, N. \end{aligned}$$

As such function may not exist for all $i = 1, \dots, N$, the *slack variables* ξ, ξ' are introduced for each point allowing for a "soft margin", and C defines the penalty imposed on observations that lie outside the ϵ margin. Compared to a Q-learning with linear FA, SVR and ERTs (the complete procedure with a pseudo-code is given in the Appendix of Geurts *et al.*, 2006) proposed by Zhao *et al.* (2009) are capable of obtaining promising results without much computational burden even in a high-dimensional data set.

As an alternative FA strategy, *deep neural networks* (DNNs; Liu *et al.*, 2017; Raghu *et al.*, 2017b,a; Atan *et al.*, 2018) has been also considered. We cover DNNs later in this section. Now we describe a completely different approach that may offer robustness to such Q-function misspecification and involves *advantage functions* rather that Q-functions.

A-learning with function approximation. A-learning (Robins, 2004; Murphy, 2003; Blatt *et al.*, 2004), where "A" stands for the "advantage" in response incurred if the optimal treatment were given relative to that actually received, is a general term used to describe a class of alternative RL methods to Q-learning, predicated on the fact that the entire Q-function needs not to be specified to estimate the optimal regime. Models can be posited only for parts of the expectation involving contrasts among treatments, as opposed to modeling the conditional expectation itself as in Q-learning. Recalling that $\mathbf{d}^* \doteq \{d_t^*\}_{t=0,\dots,T}$ denotes the optimal DTR, and denoting with $\underline{\mathbf{d}}_t^* \doteq \{d_\tau^*\}_{\tau=t,\dots,T}$ the optimal regimen from t onwards, $\mathbf{d}^{\text{ref}} \doteq \{d_t^{\text{ref}}\}_{t=0,\dots,T}$ a regimen of reference we want to make comparisons with, and with 0 the "zero-

treatment" (standard or placebo), popular contrast examples include:

$$g\left(\mathbb{E}[Y_{t+1}^{\mathbf{a_t-1},a_t,\underline{\mathbf{d}}^*_{t+1}}|\mathbf{H_t}=\mathbf{h_t}]\right)-g\left(\mathbb{E}[Y_{t+1}^{\mathbf{a_t-1},d_t^{\text{ref}},\underline{\mathbf{d}}^*_{t+1}}|\mathbf{H_t}=\mathbf{h_t}]\right), \tag{3.6}$$

$$g\left(\mathbb{E}[Y_{t+1}^{\mathbf{a_t-1},a_t,\underline{\mathbf{d}}^*_{t+1}}|\mathbf{H_t}=\mathbf{h_t}]\right)-g\left(\mathbb{E}[Y_{t+1}^{\mathbf{a_t-1},0,\underline{\mathbf{d}}^*_{t+1}}|\mathbf{H_t}=\mathbf{h_t}]\right), \tag{3.7}$$

$$g\left(\mathbb{E}[Y_{t+1}^{\mathbf{a_t-1},a_t,\underline{\mathbf{d}}^*_{t+1}}|\mathbf{H_t}=\mathbf{h_t}]\right)-g\left(\mathbb{E}[Y_{t+1}^{\mathbf{a_t-1},d_t^*,\underline{\mathbf{d}}^*_{t+1}}|\mathbf{H_t}=\mathbf{h_t}]\right), \tag{3.8}$$

where $g(\cdot)$ is a known *link function*, typically taken to be the identity link. *Optimal blip-to-reference* in (3.6) and *optimal blip-to-zero* in (3.7) evaluate removal of an amount ("blip") of treatment at stage t on the subsequent average outcome, when optimal treatment regime $\mathbf{d}^*_{t+1}$ is followed from $t+1$ onwards: "blips" are represented by the treatment of reference d_t^{ref} and the "zero-treatment", respectively. The *regret* function in (3.8) evaluates the increase in the benefit-to-go that we forego by making decision a_t rather than the optimal decision d_t^* at time t.

While Robins (2004) advocates optimal blip functions and Murphy (2003) regrets, one can notice that they are mathematically equivalent (Moodie *et al.*, 2007). In addition, they both proposed to use SNMMs for modelling the contrasts. Indeed, SNMMs represent popular models for parameterizing the conditional intermediate causal effects, that is the difference between the conditional expectation of an outcome in the observed data and the conditional expectation of an outcome under some potential outcome scenario. For instance, considering the optimal blip-to-reference function, at each $t = 0, 1, \ldots, T$, Robins (2004) proposed to generally model the causal effect as

$$g(\mathbb{E}_{\boldsymbol{d}}\left[Y_{t+1}^{\mathbf{a_t-1},a_t,\underline{\mathbf{d}}^*_{t+1}}|\mathbf{H_t}=\mathbf{h_t}\right])-g(\mathbb{E}_{\boldsymbol{d}}\left[Y_{t+1}^{\mathbf{a_t-1},d_t^{\text{ref}},\underline{\mathbf{d}}^*_{t+1}}|\mathbf{H_t}=\mathbf{h_t}\right])=\gamma_t(\mathbf{h_t},a_t;\psi_t),$$

where $g(\cdot)$ is the link function and $\gamma_t(\mathbf{h_t}, a_t; \psi_t)$ is a known $(T-t+1)$-dimensional function, smooth in ψ_t. For all $\mathbf{h_t}$, a_t, the parameterization requires $\gamma_t(\mathbf{h_t}, 0; \psi_t) = 0$, and typically, is chosen to be such that $\gamma_t(\mathbf{h_t}, a_t; 0) = 0$, so that $\psi_t = 0$ encodes the null hypothesis of no treatment effect. However, it is important to note that, while the model formulation is equivalent, the estimation technique differs. In Robins (2004), an optimal DTR, under some assumptions (see e.g., Chakraborty & Moodie, 2013), is estimated through backward recursive G-estimation; in Murphy (2003) a technique known as *iterative minimization of regrets* (IMOR) is proposed.

To provide the reader with a unified overview, making at the same time a clear distinction between the two different procedures, we will consider the terms *contrast-based A-learning* and *regret-based A-learning* to refer to Robins (2004)'s and to Murphy (2003)'s works, respectively. This is done in line with the work of Schulte *et al.* (2014), which will be taken as a guide for discussing more in depth A-learning and compare it with the widely used technique of Q-learning.

Contrast-based A-learning - We define the *optimal contrast-* or *C-function* $C_t^*(\mathbf{H_t}, A_t)$ at time t, as the expected difference in potential outcomes when using a reference regime d_t^{ref} instead of a_t at time t, and subsequently receive the optimal regime $\underline{\mathbf{d}}^*_{t+1} \doteq \{d_\tau^*\}_{\tau=t+1,\ldots,T}$. It is basically the optimal blip-to-reference given in

(3.6) with g the identity function, and it can also be expressed as a function of the optimal Q-functions as follows

$$\begin{aligned} C_t^*(\mathbf{H_t}, A_t) &\doteq \mathbb{E}(Y(\mathbf{A_{t-1}}, A_t, \underline{\mathbf{d}}_{t+1}^*) - Y(\mathbf{A_{t-1}}, d_t^{\text{ref}}, \underline{\mathbf{d}}_{t+1}^*)|\mathbf{H_t}) \\ &= Q_t^*(\mathbf{H_t}, A_t) - Q_t^*(\mathbf{H_t}, A_t = d_t^{\text{ref}}), \quad \forall t \in [0, T]. \end{aligned}$$

For simplicity, we consider here only the case of two treatment options coded as 0 and 1, i.e., $\mathcal{A}_t = \{0, 1\}$ for all $t \in [0, T]$, and we let the standard or placebo "zero-treatment" to be the reference treatment, i.e., $d_t^{\text{ref}} = 0$, leading to an equivalence between (3.6) and (3.7). To determine an optimal DTR, we begin by defining an approximation space for the contrast functions, e.g., $\mathcal{C}_t \doteq \{C_t(\mathbf{h_t}, a_t; \psi_t) : \psi_t \in \Psi_t\}$, with $\psi \in \Psi_t$, a subset of the Euclidean space. Then, in a backward fashion, starting from $t = T$, and denoting the propensity of receiving treatment $A_T = 1$ in the observed data with $\pi_T(A_T|\mathbf{h_T}) = \mathbb{P}(A_T = 1|\mathbf{H_T} = \mathbf{h_T})$, we obtain a consistent and asymptotically normal estimator for ψ_T by G-estimation (Robins, 2004), i.e., by solving estimating equations of the form:

$$\begin{aligned} 0 = \sum_{i=1}^{N} &\lambda_T(\mathbf{H_{Ti}}, A_{Ti})[A_{Ti} - \pi_T(A_{Ti}|\mathbf{H_{Ti}})] \\ &\times [Y_{(T+1)i} - A_{Ti} C_T^*(\mathbf{H_{ti}}, A_{Ti}; \psi_T) - \theta(\mathbf{H_{Ti}}, A_{Ti})], \end{aligned} \tag{3.9}$$

for arbitrary functions $\lambda_T(\mathbf{H_T}, A_{Ti})$ of the same dimension as ψ_T and arbitrary functions $\theta_T(\mathbf{H_T}, A_{Ti})$. To implement estimation of ψ_T via (3.9), one may adopt parametric models for all the unknown functions, including $\pi_T(A_{Ti}|\mathbf{H_{Ti}})$ if randomization probabilities are not known, i.e., in observational studies. Under certain conditions, Schulte *et al.* (2014) report that an optimal choice for $\lambda_T(\mathbf{H_{Ti}}, A_{Ti}; \psi_T)$ is given by $\partial/\partial\psi_T C_T^*(\mathbf{H_{ti}}, A_{Ti}; \psi_T)$. Once we get estimates $\hat{\psi}_T$, the contrast based A-learning algorithm iteratively proceeds by estimating $\hat{\psi}_{T-1}, \hat{\psi}_{T-2}, \ldots, \hat{\psi}_0$. Finally, in this two treatment setting, the optimal DTRs is given by the one which positives the C-function, i.e.,

$$\hat{d}_t^*(\mathbf{H_t}) = d_t^*(\mathbf{H_t}; \hat{\psi}_t) = \mathbb{I}(C_t^*(\mathbf{H_t}, A_t; \hat{\psi}_t) > 0), \quad \forall t \in [0, T].$$

Notice that, as the additional models specified in (3.9) are only adjuncts to estimating ψ_T, as long as at least one of these models is correctly specified, (3.9) will provide a consistent estimator for ψ_T (this property is called *double robustness property*). In contrast, Q-learning requires correct specification of all Q-functions. An intermediate approach between G-estimation and Q-learning, which affords double-robustness to model misspecification and requires less computational skills compared to the former, was later introduced by Wallace & Moodie (2015) as the *dynamic weighted ordinary least squares* (dWOLS).

Regret based A-learning - Rather than modelling a contrast defined as the expected difference in outcome when using a reference regime d_t^{ref} instead of a_t at time t, Murphy (2003) and Blatt *et al.* (2004) proposed to model a regret function similar to the one introduced in (3.8). Denoting it by μ_t^*, it is defined as $\mu_t^*(\mathbf{H_t}, A_t) \doteq \mathbb{E}(Y(\mathbf{A_{t-1}}, A_t, \underline{\mathbf{d}}_{t+1}^*) - Y(\mathbf{A_{t-1}}, d_t^*, \underline{\mathbf{d}}_{t+1}^*)|\mathbf{H_t})$, for $t \in [0, T]$. Here the "advantage"/regret, is the gain/loss in performance obtained by following action A_t at

time t and thereafter the optimal regime $\underline{\mathbf{d}}^*_{t+1}$ as compared to following the optimal policy $\underline{\mathbf{d}}^*_t$ from time t on. Again, to estimate the optimal treatment regime, we model the regrets by defining an approximation space for the t-th advantage μ-function, e.g. $\mathcal{M}_t \doteq \{\mu_t(\mathbf{h_t}, a_t; \psi_t) : \psi_t \in \Psi_t\}$, with $\psi \in \Psi_t$, a subset of the Euclidean space. As with Q-learning and in contrast-based A-learning, we use ADP and permit the estimator to have different parameters for each time t, but in this case an alternative estimation strategy, known as IMOR, was proposed Murphy (2003). It is based on simultaneously estimating the regret model parameter ψ plus a c parameter used for improving the stability of the algorithm, by searching for $(\hat{\psi}, \hat{c})$ that satisfy

$$\sum_{t=0}^{T} \mathbb{P}_N \left[Y + \hat{c} + \sum_{\tau=0}^{T} \mu_\tau(\mathbf{H}_\tau, A_\tau; \hat{\psi}) - \sum_a \mu_t(\mathbf{H_t}, a; \hat{\psi}) \pi_t(a|\mathbf{H_t}; \hat{\alpha}) \right]^2$$
$$\leq \sum_{t=0}^{T} \mathbb{P}_N \Big[Y + c + \sum_{\tau \neq t} \mu_\tau(\mathbf{H}_\tau, A_\tau; \hat{\psi}) + \mu_t(\mathbf{H_t}, A_t; \psi)$$
$$\sum_a \mu_t(\mathbf{H_t}, a; \psi) \pi_t(a|\mathbf{H_t}; \alpha) \Big]^2,$$

for all ψ and c, with $\mathbb{P}_N$ denoting the empirical mean of a sample of N patients. This technique iteratively searchs a solution until convergence. We point to the original work of Murphy (2003) for readers interested in this technique and its relationship with G-estimation.

Q-learning and A-learning are probably the two most widely used approaches for optimal DTRs estimation. They have been extensively studied and discussed in literature, and I want to report here some of the considerations and results related to the study of the trade-offs between the two off-line RL-based techniques.

First, Chakraborty *et al.* (2010) showed how, under some sufficient conditions, Q-learning with linear models is algebraically equivalent to an inefficient version of G-estimation with regret functions. The reasoning is based on the following relationship between the contrast and Q-functions:

$$\mu^*_t(\mathbf{H_t}, A_t) \doteq \mathbb{E}(Y(\mathbf{A_{t-1}}, A_t, \underline{\mathbf{d}}^*_{t+1}) - Y(\mathbf{A_{t-1}}, d^*_t, \underline{\mathbf{d}}^*_{t+1}) | \mathbf{H_t}) =$$
$$= Q^*_t(\mathbf{H_t}, A_t) - \max_{a_t \in \mathcal{A}_t} Q^*_t(\mathbf{H_t}, A_t), \quad \forall t \in [0, T].$$

Second, Schulte *et al.* (2014) evaluated through MC simulations their finite sample performance in different scenarios, aiming at identifying regions in which one method is superior to the other, and provided a detailed and self-contained comparison of A-learning and Q-learning. In a two-stages ($T = 1$) experiment with two treatment options, they found that when all models were correctly specified, Q-learning was markedly more efficient (nearly twice) in estimating the second stage ($t = 1$) parameters, but only modestly so for first stage ($t = 0$) parameters. Q-learning was also more efficient than A-learning when the propensity model was misspecified. However, if the Q-function was misspecified, there were values of the parameters for which gains in efficiency exhibited by Q-learning were clearly outweighed by the bias incurred, making A-learning preferable in terms of mean squared error. Finally, with both propensity model and Q-learning models misspecified there was

no general trend in efficiency of estimation across parameters that might recommend one method over the other.

Finally, Moodie *et al.* (2007) also discussed the relationship between the different proposed strategies, focusing on the similarities between IMOR and G-estimation. They show how the IMOR optimization technique of Murphy (2003) is a special case of G-estimation (Robins, 2004) under the null hypothesis of no treatment effect, and modeling by a constant.

Deep Reinforcement Learning. The tremendous success achieved in recent years by RL in many complex domains has been largely enabled by the use of advanced FA techniques in combination with large-scale data generation, particularly from self-play games (Jonsson, 2019). As seen for Q-learning, in most realistic settings, when the state space $\mathcal{X}$ is too large, it is common to parameterize the Q-function (or the value function or the policy itself) on some parameter vector $\boldsymbol{\theta}$. The value in a state is then completely determined by the current parameters estimates, and the update rules for RL algorithms are modified such that they no longer update the values of states directly, but rather the parameters in $\boldsymbol{\theta}$. In large part, the great success achieved by RL in intelligent game playing, e.g., AlphaGo (Silver *et al.*, 2017) or Atari (Mnih *et al.*, 2015), has been made possible by powerful FA methods in the form of *deep neural networks* (DNNs).

Combining RL and DNNs for approximating these quantities gives rise to *deep RL* (DRL), which includes *deep Q-network* (DQN), or *deep Q-learning*, as an example. In a DQN, a DNN (e.g., *feed-forward*, *recurrent*, *convolutional*; see Goodfellow *et al.*, 2016, for an overview of existing architectures) is used to approximate the Q-function. More specifically, at each time t, a DNN is used to fit a model for the Q-function in a supervised way: states and actions $\{(\mathbf{H}_{t,i}, A_{t,i})\}_{i=1,\dots,N}$ are given as input, and the Q-values of all possible actions are generated as output $\{Q_t(\mathbf{H}_{t,i}, A_{t,i}; \hat{\mathbf{W}}, \hat{\mathbf{b}})\}_{i=1,\dots,N}$, leading to a labelled set of data $\mathcal{D} = \{(\mathbf{H}_{t,i}, A_{t,i}), Q_t(\mathbf{H}_{t,i}, A_{t,i}; \hat{\mathbf{W}}, \hat{\mathbf{b}})\}_{i=1,\dots,N}$. $\mathbf{W}$ and $\mathbf{b}$ represent the unknown *weight* and *bias* parameters of the DNN, respectively. Figure 3.5 shows a schematic of a *feed-forward neural network* (FFNN) used within RL. It is characterized by a set of neurons, structured in layers: the input layer, the output layer and the hidden intermediate layers. Each neuron processes the information forward from one layer to the next one through a pre-specified activation function depending on the unknown parameters. Collected data $\mathcal{D}$ is stored and continuously updated by the user in memory for updating Q-function parameters' estimates. Next action is determined by an exploration scheme (typically ϵ-greedy) which probabilistically chooses between the action with the highest Q-value as in Algorithm 1 and a random action. For updating the Q-network, we minimize a loss function, generally the MSE between our target Q-value and our current Q-output, and this is efficiently done by a technique known as back-propagation or stochastic gradient descent (LeCun *et al.*, 1998; Goodfellow *et al.*, 2016).

Within the DTR literature, DRL implementations for estimating optimal regimes, have been proposed in Liu *et al.* (2017) and Raghu *et al.* (2017b), for the graft-versus-host disease after transplantation and sepsis treatment, respectively. Both works use observational medical data and are build on the DQN developed in Mnih *et al.* (2015), for which an illustration is available in Algorithm 2. More recently, Atan *et al.* (2018) proposed a more sophisticated approach for constructing effective treatment

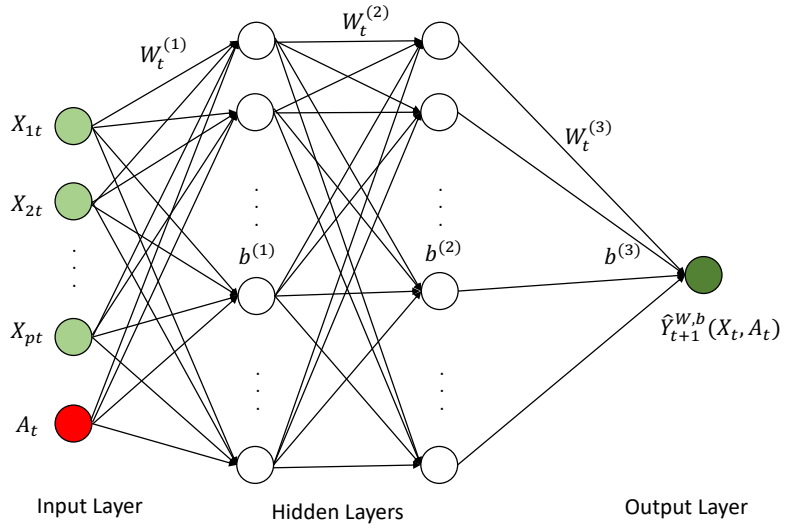

Figure 3.5. Representation of a feed-forward neural network with four layers used within RL. In the first (input) layer, we introduce our input data, covariates $X_{1t}, \ldots, X_{pt}$ and treatment A_t at time t, which are non-linearly transformed according to their weights $W^{(1)}$ and a bias parameter $b^{(1)}$ through the neurons of the first hidden layer. In a deep setting, many hidden layers may process the information, this is what make deep learning different from a simple neural network. The final (output) layer, generates the predicted outcome value (reward) $\hat{Y}_{t+1}^{\mathbf{W},\mathbf{b}}(X_t, A_t)$, with $\mathbf{W}$ and $\mathbf{b}$ representing the parameters of the deep neural network.

policies when the observed data is biased and lacks counterfactual information. Their approach separates the problem into two stages: first the bias is reduced by learning a representation map using an auto-encoder architecture for a DNN, then a FFNN is used on the transformed data to estimate an optimal DTR.

DNNs allow for model flexibility and process features without the need of domain knowledge, being particularly suitable for real-life complexity, high dimensionality, and adaptivity. Compared to their shallow counterpart, they are more capable of automatic feature representation and capturing complicated relationships. However, one general limitation of indirect methods such as Q- and A-learning, independently on the FA, is that the optimal DTRs are estimated in a two-step procedure: one estimates either the Q-functions or the contrast/regret functions using the data; then these functions are either maximized or minimized to infer the optimal DTR. In the presence of high-dimensional information, even with flexible non-parametric techniques such as SVR and DL, it is possible that these conditional functions are poorly fitted, and thus the derived DTR may be far from optimal. Moreover, this approach may not necessarily result in maximal long-term clinical benefit, as demonstrated by Zhao *et al.* (2012) who shifted to parameterize and estimate the treatment rule directly.

Bayesian approaches. Many Bayesian methods have been used in practice to identify optimal DTRs (Thall *et al.*, 2007; Arjas & Saarela, 2010; Zajonc, 2012; Xu *et al.*, 2016). However, they are likelihood-based methods, requiring thus a joint estimation of data trajectory, and then either apply DP or a full numerical search of the action space to identify the optimal DTR. In order to bridge the gap between

Algorithm 2: Deep Q-Network (Mnih *et al.*, 2015; Liu *et al.*, 2017)

Input: Pre-processing real data profiles $\{(\mathbf{H}_{0,i}, A_{0,i}), Y(\mathbf{H}_{0,i}, A_i)\}_{i=1,\dots,N}$; $\epsilon > 0$; $\gamma > 0$.

Initialization: Experience memory $\mathcal{D}_0 = \{(\mathbf{H}_{0,i}, A_{0,i}), Q_0(\mathbf{H}_{0,i}, A_{0,i}; \mathbf{W}_0, \mathbf{b}_0)\}_{i=1,\dots,N}$ with $\{Q_0(\mathbf{H}_{0,i}, A_{0,i}; \mathbf{W}_0, \mathbf{b}_0)\}_{i=1,\dots,N}$ based on random parameters $(\mathbf{W}_0, \mathbf{b}_0)$.

Train a DNN with labelled data $\mathcal{D}_0$ and get estimates $(\hat{\mathbf{W}}_0, \hat{\mathbf{b}}_0)$

for $t = 0, 1, 2, \dots T$ **do**

- **ϵ-Greedy step**: select a random action a_t with probability ϵ; otherwise $a_t = \arg\max_{a \in \mathcal{A}} Q_t(\mathbf{H}_t, a; \hat{\mathbf{W}}_t, \hat{\mathbf{b}}_t)$;
- Execute a_t in emulator and observe state transition X_{t+1} and reward $Y_{t+1}(\mathbf{H}_t, a_t)$;
- Update the experience memory data: $\mathcal{D}_{t+1} = (\mathcal{D}_t, Y_{t+1}, X_{t+1})$;
- **Q-learning Update**: update Q-function

$$Q_t(\mathbf{H}_t, a_t) = Y_{t+1}(\mathbf{H}_t, a_t) + \gamma \max_a Q_{t+1}(\mathbf{H}_{t+1}, a)$$

- **DNN Update**: get updated estimates $(\hat{\mathbf{W}}_{t+1}, \hat{\mathbf{b}}_{t+1})$ with SGD for minimizing the expected loss $\left(Q_t(\mathbf{H}_t, a_t) - Q_t(\mathbf{H}_t, a; \hat{\mathbf{W}}_t, \hat{\mathbf{b}}_t)\right)^2$ based on the Q-learning update.

end for

Bayesian inference and existing RL-based DTRs approaches, Murray *et al.* (2018) first proposed the so called *Bayesian Machine Learning* (BML), which allows both for patient's preferences and physician's expert knowledge to be incorporated in the model, and the flexibility of novel ADP approach. The BML proposal fits a series of Bayesian regression models (authors recommend using Bayesian non-parametric regression models), one for each stage, in reverse sequential order. One distinguishing feature of BML is that it treats the counterfactual response variables as missing values, and multiply imputes them from their posterior predictive distribution, which is derived from the previously fitted regression models. A detailed presentation of Bayesian methodologies and the many modeling choices required for a Bayesian estimation of a DTR is beyond the scope of this text, however a great number of resources are available to the interested reader (e.g., Chen *et al.*, 2010).

Direct RL methods

Direct methods, also known in RL literature as direct policy search methods (Ng *et al.*, 2000), seek to maximize the return (i.e., the discounted sum of future rewards, see Section 2) by learning the optimal policy or value directly, without involving intermediate quantities such as Q-functions or C-functions. These methods typically do not assume models for conditional means or other aspects of the conditional distributions of the the outcomes; in this sense they are called "non-parametric". However, they may consider a parametrization of the class of policies.

In direct methods, indeed, first a class of policies or regimes $\mathcal{D}$, often indexed by a parameter, say $\psi \in \Psi$, is pre-specified. Then, for each candidate regime $d \in \mathcal{D}$, an estimate $\hat{V}_d = \hat{V}_d(X_0)$ of the corresponding value is obtained. Recall from Section

2 that the value is the mean of the return marginalized over all observations that might be impacted by the treatment (see 2.5). The regime that maximizes this value function represents the optimal treatment regime d^*

$$\hat{d}^* \doteq \arg\max_{d \in \mathcal{D}} \hat{V}_d = \arg\max_{\psi \in \Psi} \hat{V}_{d_\psi}. \tag{3.10}$$

For a simple example of a parametric class of policies, consider DTRs that use a suitable summary of the available history (*tailoring variable*) to indicate when to change treatment: if the *tailoring variable* falls below/above a threshold ψ, treatment is changed. Another common example is given by the *soft-max* class of functions $\mathcal{D} \doteq \{\pi(a_k|\mathbf{x}, \psi) = e^{-\mathbf{x}^T\psi_k} / \sum_{j=1}^K e^{-\mathbf{x}^T\psi_j} : \psi \in \Psi, k = 1, \dots, K\}$, where $a_1, \dots, a_K$ denote the K possible treatments and $\psi \doteq (\psi_1^T, \dots, \psi_K^T)$ the vector of parameters for the K treatments indexing the class of policies.

Most of the statistical work in this area is based on the IPTW technique (Robins *et al.*, 1994). It is used, for instance, in estimating MSMs (Robins, 2000; Orellana *et al.*, 2010) or value functions (Zhang *et al.*, 2012b, 2013); in classification-based frameworks, such us *outcome weighted learning* (OWL; Zhao *et al.*, 2012, 2015; Liu *et al.*, 2018), and in combination with ML approaches, such as decision trees (Laber & Zhao, 2015; Tao *et al.*, 2018).

Inverse Probability of Treatment Weighting. IPTW is a general technique that can be used in DTRs for inferring causal effects from observational data, under the standard assumptions for causal inference discussed in Section 3.1.1.

In case of a randomized trial, particularly SMART designs (see Section 3.1), estimating an optimal regime based on 3.10 is relatively straightforward. The *target policy* we learn $\mathbf{d}$ corresponds to the fixed *exploration policy* $\boldsymbol{\pi}$ that generated the trajectories: it consists in the randomization probabilities and is known by design. On the contrary, when this information is not available, as in the case of the most common observational studies, the value function has to be estimated for an arbitrary treatment policy $\mathbf{d}$ using an empirical sample of N trajectories. This learning procedure is also called *off-policy (batch) learning*. Making use of the importance sampling technique, which assumed P_d absolutely continuous with respect to P_π (corresponding to the positivity assumption in 3.1), we change the distribution under which we compute the value function. In doing that, we basically weight our returns according to the relative probability of their trajectories occurring under the target and exploration policies:

$$\begin{aligned} V_d = \mathbb{E}_d[Y] = \int Y dP_d = \int Y \left(\frac{dP_d}{dP_\pi}\right) dP_\pi = \\ = \int \left(\prod_{t=0}^{T} \frac{\mathbb{I}[A_t = d_t(H_t)]}{\pi_t(A_t|H_t)}\right) Y dP_\pi \doteq \int w_{d,\pi} Y dP_\pi. \end{aligned} \tag{3.11}$$

Now, a natural way to estimate V_d is given by its Monte Carlo (MC) estimator, i.e., $\hat{V}_d \doteq \mathbb{P}_N [w_{d,\pi} Y]$, where $\mathbb{P}_N$ denotes the empirical average over N patients' trajectories. The MC estimator is known to be an unbiased estimator (by the *Strong Low of Large Numbers*), but its variance is unbounded. To this purpose, to obtain a more stable estimator, the weights $w_{d,\pi}$ are normalized by their sample mean,

leading to the IPTW estimator (Robins, 2000)

$$\hat{V}_d^{IPTW} \doteq \frac{\mathbb{P}_N\left[w_{d,\pi}Y\right]}{\mathbb{P}_N\left[w_{d,\pi}\right]}. \tag{3.12}$$

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

When π is known (e.g., SMART design), the IPW estimator is consistent. However, it is highly variable due to the presence of the non-smooth indicator functions inside the weights.

An alternative version, which integrates the properties of the IPTW estimator with those of the regression based estimator, assuming models for both the propensity score and the (conditional) mean outcome, is the *augmented inverse probability of treatment weighting* (AIPTW) estimator (Zhang *et al.*, 2012b). Its original version, was designed for a single stage treatment regime, thus, does not make use of any RL strategies. Assuming a single-stage treatment regime with two treatment options ($A \in \{a, a'\}$), let $H = X_0$ denote patient's history, $d(H) \doteq d(H;\psi)$ a treatment regime indexed by ψ, $\mu(A, H; \hat{\beta})$ an estimated model for the mean outcome $\mathbb{E}[Y|H, A]$, and $\pi(A|H, \hat{\gamma})$ an estimated propensity score. Then, the AIPW estimator is defined by

$$\hat{V}_d^{AIPW} \doteq \mathbb{P}_N\left\{\frac{\mathbb{I}[A = d(H;\psi)]Y}{\pi(H;\psi,\hat{\gamma})} - \frac{\mathbb{I}[A = d(H;\psi)] - \pi(H;\psi,\hat{\gamma})}{\pi(H;\psi,\hat{\gamma})}\mu(H;\psi,\hat{\beta})\right\}, \tag{3.13}$$

where,

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

It only requires either the propensity or mean outcome model to be correctly specified but not both, hence, *doubly robust* method. In addition to being more robust to model mis-specification, AIPW estimators tend to be more efficient than their non-augmented counterparts (Robins, 2004).

An adaptation of this approach from the single-stage setting to sequential treatment decisions is available in Zhang *et al.* (2013); Tao & Wang (2017); Zhang & Zhang (2018), where, with models posited for either Q-functions or C-functions, a Q-learning or Contrast-based A-learning strategy was combined with the IPTW estimation, making it fully RL based.

IPTW represents a basis for other existing direct methods. For instance, it constitutes one of the most common approach for estimating MSMs, introduced in the causal inferences literature for controlling for confounding through assigning each participant a weight (Robins, 2000; Neugebauer *et al.*, 2012). MSMs are a powerful alternative to SNMMs for describing the causal effect of a treatment (hence "structural"), and pertain to population-average effects ("marginal" over covariates, baseline and time-varying outcomes). They basically represent models for the expectation of a potential outcome under a specified unobserved DTR $\mathbf{d}$,

marginalizing over the covariate history $V_{\mathbf{d}} = \mathbb{E}_{\mathbf{d}}[Y] = \mathbb{E}[Y^{\mathbf{d}}]$, or alternatively as a function of the baseline covariates X_0 only, i.e., $V_{\mathbf{d}}(X_0) = \mathbb{E}_{\mathbf{d}}[Y|X_0] = \mathbb{E}[Y^{\mathbf{d}}|X_0]$. Examples and a further discussion is provided in Appendix A.

Other IPTW-based methods are developed within the general framework proposed by Zhang *et al.* (2012a) and Zhao *et al.* (2012), who recast the estimation of the optimal decision rule as a classification problem. We illustrate now this framework and the specific OWL approach Zhao *et al.* (2012), with some of the subsequent developments.

Outcome Weighted Learning. As an alternative direct approach, Zhao *et al.* (2012) reformulated the problem of optimal DTRs estimation as a weighted classification problem, with weights retrospectively determined from clinical outcomes (from here "Outcome Weighted Learning"); and proposed to solve it with tools of ML literature.

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

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

However, as solving the problem is hard due to the discontinuous indicator function, Zhao *et al.* (2012) proposed to address it with a convex surrogate loss function for the $0-1$ loss, which corresponds to the *hinge loss* used for SVM optimization (Hastie *et al.*, 2009). Considering that $d(H)$ can always be represented as $sign(f(H))$ for some suitable function f, the corresponding minimization problem proposed by the authors can be given as

$$\hat{f}^* \doteq \arg\min_{f\in\mathcal{F}} \mathbb{P}_N \left[\frac{Y}{\pi(A|H)} \phi(Af(H)) + \lambda_N ||f(H)||^2\right], \tag{3.14}$$

where λ_N is a tuning penalty parameter that can be chosen via cross-validation, and $\phi(x) \doteq \max(1 - x, 0)$ is the hinge loss.

Although the seminal work of Zhao *et al.* (2012) allows the use of different loss functions, the specific considered setting (non-negative rewards, single stage, binary treatments) opened many problems for its practical employment, some of which have been addressed by subsequent DTR literature. We report them and illustrate their relative progress with respect to the basic OWL estimator in (3.14) in **Table** 3.2.

However, note that while a broad literature focused on the single stage setting, only two works (Zhao *et al.*, 2015; Liu *et al.*, 2018) proposed an extension to multiple stages, integrating the OWL estimator and the RL framework. Zhao *et al.* (2015) developed the so-called *Backward Outcome Weighted Learning* (BOWL) and *Simultaneous Outcome Weighted Learning* (SOWL) procedures. In the first approach, the stage-t estimator which we denote with $\hat{f}^*_{B,t}$ is obtained recursively by

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

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

Reference & Technique	Extension from the standard OWL (Zhao *et al.*, 2012)
Zhao *et al.* (2015) BOWL + SOWL	**From a single stage to general T-stages**, with $T < \infty$. Authors proposed two methods: one performs an iterative backward OWL (BOWL) estimation, the other a simultaneous OWL (SOWL) estimation. Both are based on the original OWL.
Liu *et al.* (2018) AOL	**Extends to negative outcomes and considers multiple stages**. Authors proposed an augumented version for the weight of the OWL (AOL) integrating OWL and Q-functions. The robust augmentation, making use of predicted pseudo-outcomes from regression models for Q-functions, reduces the variability of weights and improves estimation accuracy.
Zhou *et al.* (2017) RWL	**Suitable for continuous, binary and count outcomes, and performs variable selection**. Authors proposed a general framework, called Residual Weighted Learning (RWL) for improving the finite sample performance, where they employ a *smoothed ramp loss* and derive outcome residuals with a regression model.
Chen *et al.* (2018) GOWL	**Suitable for ordinal treatments and negative outcomes**. Authors generalize the OWL (GOWL) by using a modified loss function and a reformulation of the objective function in (3.14).
Zhang *et al.* (2020a) MOML	**Extension to multicategory treatment scenarios and negative outcomes**. Authors use sequential binary methods by proposing a margin-based learning (build upon the *large-margin unified machine loss*), which has a special case the standard OWL.
Fu *et al.* (2019) ROWL	**Tackles the problem of outliers and considers multicategory treatments and negative outcomes**. Authors propose a robust OWL (ROWL), based on an angle-based classification structure, designed for multicategory classification problems, and a new family of *robust loss* functions to build more stable DTRs.

Table 3.2. Statistical methods which extended the standard OWL estimator in (3.14) for developing optimal treatment regimes.

For the second approach, a simultaneous estimation is performed with

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

where $\psi(x_1, x_2) \doteq \min(x_1 - 1, x_2 - 1, 0) + 1$ is a concave surrogate for the product of two indication functions.

Even if in numerical examples both BOWL and SOWL have demonstrated superior performances to existing direct methods, significant information loss is registered as t decreases. To overcome this problem, an augumented version integrating OWL and Q-functions is proposed in Liu *et al.* (2018). Defining and denoting a *pseudo-outcome* with $\tilde{Y}_t \doteq Y_t + \hat{Q}_{t+1} - \hat{s}_t(\mathbf{H_t})$, this is given by

$$\hat{f}_{A,t}^* \doteq \arg\min_{f\in\mathcal{F}} \mathbb{P}_N \left[\frac{|\tilde{Y}_t|}{\pi_t(A_t|\mathbf{H}_t)} \phi(A_t sign(\tilde{Y}_t) f_t(\mathbf{H}_t)) + \lambda_N ||f_t(\mathbf{H}_t)||^2 \right],$$

where $\hat{s}_t(\mathbf{H_t})$ is estimated via a least squares regression that minimizes $\mathbb{P}_N[Y_t + \hat{Q}_{t+1} - s_t(\mathbf{H_t})]^2$, and function ϕ is analogous to the one previously defined.

Tree-based methods. Again, by integrating tools from the ML literature, first Laber & Zhao (2015), in the context of individualized (single stage) treatment regimes, and then Tao *et al.* (2018) and Sun & Wang (2020) for dynamic regimes, introduced the tree-based approach (Breiman *et al.*, 2001) for directly estimating optimal DTRs. The underlying idea of Tao *et al.* (2018) is, first, to define and estimate a *purity*, i.e., a target measure or output which needs to be optimized, and then, to improve the purity with a decision tree. Improvement is performed by splitting a *parent node* into *child nodes* repeatedly, and by choosing a split among all possible splits at each node so that the resulting child nodes are the purest (e.g., having the lowest misclassification rate). The mean outcome (or value function), is used as purity measure, and its estimation is carried out with the IPTW estimator (Robins, 2000), or alternatively a kernel smoother in the case of continuous treatments (Laber & Zhao, 2015), and the AIPTW estimator Zhang *et al.* (2012b), respectively. Differently, Sun & Wang (2020) proposed a stochastic tree-based reinforcement learning which uses Bayesian additive regression trees, and then stochastically constructs an optimal regime using a Markov chain Monte Carlo (MCMC) tree search algorithm. In the multiple stages setting, estimation is implemented recursively using backward induction, starting from $t = T + 1$ and using the outcome Y_{T+1} directly.

By combining the properties of a tree-based learning (straightforward to understand and interpret, and capable of handling various types of data without distributional assumptions) with those of the AIPTW (semi-parametric robust estimator), the tree-based approaches are robust, efficient and more interpretable and flexible (compared to the OWL, or the DNN, for instance) in the identification of optimal DTRs.

RL Methods for Indefinite-Horizon Problems

While in computer science, a vast literature on estimating optimal policies over an increasing time horizon exists (Szepesvári, 2010; Sugiyama, 2015), this scenario

is rare in DTR literature. By adopting backward induction, most of the existing methods cannot extrapolate beyond the time horizon in the observed data. However, for some chronic conditions, or those with very short time steps, including mHealth JITAIs (see Section 3.1), the time horizon is not definite, in the sense that treatment decisions are made continually throughout the life of the patient, with no fixed time point for the final treatment decision.

To the best of our knowledge, only two proposals (Ertefaie & Strawderman, 2018; Luckett *et al.*, 2020) have been advanced in DTR literature for indefinite-horizon tasks, one of which, i.e., Luckett *et al.* (2020), motivated by a real mHealth study for treating patients with diabetes over the long term. We now review these methods; they are both developed under a time-homogeneous Markov behaviour, and, while the *V-learning* technique of Luckett *et al.* (2020) directly maximises the policy (direct-RL), the alternative *Greedy Gradient Q-learning (GGQ)* of Ertefaie & Strawderman (2018) uses indirect RL.

Greedy Gradient Q-learning (GGQ). The first extension of DTRs estimation in indefinite-horizon problem was introduced by Ertefaie & Strawderman (2018). Motivated by the original GGQ algorithm of Maei *et al.* (2010), they proposed a generalization of the GGQ imposing a time-homogeneous Markov assumption (see Section 2.0.1) on the state-action sequences for each subject. Although not imposed by other DTRs methods, this assumption exemplifies estimation and inference by working with time-independent Q-functions and optimal regimes, and avoiding the need for backward induction, which has time horizon limitations.

We adopt similar notation as in the previous sections, with the introduction of an absorbing state c representing, for instance, a death event. We assume that at each time t patients' covariates X_t take values in the state space $\mathcal{X}^\star \doteq \mathcal{X} \cup \{c\}$, with $\mathcal{X} \cap \{c\} = \emptyset$. We remind that in time-homogeneous MDPs, transition probabilities $\{p_t\}_{t\geq 1}$, states and actions spaces are time-independent. Let also the state and action spaces be finite, with the action space $\mathcal{A}_x$ defined by the covariates' information. $\mathcal{A}_x$ consists of $0 < m_x \leq m$ treatments, with m the total number of available treatments during all the steps. For any t such that $X_t = c$, let $A_x = \mathcal{A}_c = \{u\}$, where u denotes "undefined"; this implies that $p(X_{t+1} = c, A_{t+1} = u | X_t = c, A_t = y) = 1$.

Now, denoted with $\widetilde{T} \doteq \inf\{t > 0 : X_t = c\}$ a stopping time (death for example), individual trajectories, including also the last final state, will be given by $(X_0, A_0, R_1, \ldots, X_{\widetilde{T}-1}, A_{\widetilde{T}-1}, R_{\widetilde{T}}, X_{\widetilde{T}})$. Note that $\mathbb{P}(\widetilde{T} < \infty | X_0, A_0) = 1$, regardless of (X_0, A_0).

Based on these specifications, under the causal inference assumptions, one can define the infinite time-horizon stage t action-value function for a specified deterministic regime $\pi(\mathbf{h_t}) = \pi(x_t) = \pi(x)$, for $x \in \mathcal{X}$, as

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

From Section 2, we know that the optimal Q-function satisfies the Bellman optimality equation given in (2.12). In addition, we also have that $Q^*(c,a) = 0$, as the return is set to 0 after a patients is lost to follow-up.

For estimating an optimal regime, Ertefaie & Strawderman (2018) proposed to estimate the optimal Q-function $Q^*(x,a)$ with linear FA (as illustrated in Section

3.1.1). Let $Q(x,a;\theta^*)$ be a parametric model for $Q^*(x,a)$ indexed by $\theta^* \in \Theta \subseteq \mathbb{R}^q$, and suppose a linear model with interactions, i.e. $Q(x,a;\theta^*) = \theta^{*T}\psi(x,a)$, with $\psi(x,a)$ being a known q-dimensional vector of features summarizing the state and treatment pair. To ensure $Q^*(c,a) = 0$, we also need $\psi(c,a) = 0$.

Now, the Bellman optimality equation suggests and motivates the following unbiased estimating function for θ^*

$$\hat{D}(\theta^*) = \mathbb{P}_N \left\{ \sum_{t=0}^{T-1} \left(Y_{t+1} + \gamma \max_{a' \in \mathcal{A}_{X_{t+1}}} Q(X_{t+1}, a'; \theta^*) - Q(X_t, A_t; \theta^*) \right) \psi(X_t, A_t) \right\}, \tag{3.15}$$

where $\mathbb{P}_N$ denote the empirical average on N i.i.d. trajectories, and $\psi(X_t, A_t) \doteq \nabla_{\theta^*} Q(X_t, A_t; \theta^*)$.

However, the estimating function in (3.15) is a continuous, piecewise-linear function in θ^* that is non-convex and non-differentiable everywhere. To overcome the problem, under some regularity conditions, any solution $\hat{\theta}^*$ is equivalently defined as a minimizer of $\hat{M}(\theta^*) \doteq \hat{D}(\theta^*)^T \hat{W}^{-1} \hat{D}(\theta^*)$, with $\hat{W} \doteq \mathbb{P}_N \left\{ \sum_{t=0}^{T-1} \psi(X_t, A_t)^{\otimes 2} \right\}$, and $x^{\otimes 2} \doteq xx^T$, for any vector x (Ertefaie & Strawderman, 2018). If $\hat{\theta}^* = \arg\min_{\theta^* \in \Theta} \hat{M}(\theta^*)$ is the unique solution, then $\hat{Q}^*(x,a) = Q(x,a;\hat{\theta}^*)$ and the corresponding optimal regime is given by $\hat{\pi}^* = \arg\max_{a \in \mathcal{A}_x} Q(x,a;\hat{\theta}^*)$.

GGQ's performance has been demonstrated in the context of chronic diseases with large sample sizes and a moderate number of time points. Under additional assumptions, authors also prove that $\hat{\theta}^*$ is a consistent estimator for θ^* and asymptotically normally distributed.

V-learning . In the GGQ method of Ertefaie & Strawderman (2018), the estimated policy is based on the estimating equation in (3.15), which contains a non-smooth max operator that makes estimation difficult without large amounts of data (Laber *et al.*, 2014b; Linn *et al.*, 2017), and, depending directly on $\hat{\theta}^*$, it requires modeling the transition probabilities. Motivated by a mHealth application, where policy estimation is continuously updated in real time as data accumulate (starting with small sample sizes), an alternative method, which directly maximizes estimated values over a class of policies, was proposed in Luckett *et al.* (2020).

Under the same causal inference and time-homogeneous MDP assumptions of Ertefaie & Strawderman (2018), and provided interchange of the sum and integration is justified, the targeted state-value function of policy d is state x_t is

$$V(x_t) = \sum_{\tau \geq t} \mathbb{E}\left[\gamma^{\tau - t} Y_{\tau+1} \left(\prod_{v=0}^{\tau} \frac{d(A_v|X_v)}{\pi_v(A_v|S_v)} \middle| X_t = x_t \right) \right],$$

π and exploration policy, which can be seen as the randomization probability in an RCT, and d an arbitrary policy which we want to learn about.

In light of the Bellman equation in (2.10) for the value function, it follows that, for any function ψ defined on the state space $\mathcal{X}_t$, the state-value function satisfies

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

which represents an importance-weighted variant of the Bellman optimality (Sutton & Barto, 2018).

Let now $V(x;\theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^q$, denote a model for $V(x)$. Assuming that $V(x;\theta)$ is differentiable everywhere in θ, for fixed x and d, let $\psi(x)$ be the gradient of $V(x;\theta)$, i.e., $\psi(x) \doteq \nabla_\theta V(x;\theta)$, and define the alternative estimating equation function as

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

Similarly to (3.15), $\hat{\theta}$ can be defined as the minimizer of $\hat{M}(\theta) \doteq \hat{\Lambda}(\theta)^T \hat{W}^{-1} \hat{\Lambda}(\theta) + \lambda \mathcal{P}(\theta)$, with $\hat{W}$ a positive definite matrix in $\mathbb{R}^{q\times q}$, λ a tuning parameter and $\mathcal{P} : \mathbb{R}^q \to \mathbb{R}_+$ a penalty function. Subsequently, $V(x;\hat{\theta})$ is the estimated state-value function under d in state x, and the estimated optimal regime $\hat{d}^*$ is the one that maximises the estimated value function.

V-learning only requires modeling the policy and the value function, rather than the data-generating process. In addition, by directly maximizing the estimated value over a class of policies (see Luckett *et al.*, 2020, for more details) it avoids the non-smooth max operator in (3.15). The developed RL method is applicable over indefinite horizons and is suitable for both off-line and online learning. Thus, it can be successfully applied in the novel mHealth field which we will discuss soon in Section 3.1.2.

Real-World DTR Studies using RL and Practical Challenges

In the previous sections we reviewed existing RL-based methodologies for estimating optimal DTRs. These were introduced within the ML and statistical literature and generally evaluated through simulations. Now, we want to provide a more concrete idea of applications of RL for estimating optimal DTR in real-world settings, as found in clinical literature. At the same time, we want to illustrate the main challenges that clinical researchers face in applying these methods in practical settings, and the main limitations that might impact a successful output.

Generally speaking, based on the type of the disease, we can distinguish two main domains: chronic diseases and critical care. Chronic diseases are defined broadly as long-lasting conditions (three months or more) and require ongoing medical attention or limit activities of daily living (Bernell & Howard, 2016). They include the leading causes of death and disability (e.g., cancer, cardiovascular diseases, diabetes, mental illness, obesity) and are also the main drivers of nations healthcare costs (Organization *et al.*, 2018, 2005). Typically, apposite protocols (Wagner *et al.*, 2001) support practitioners to facilitate decision making. However, such protocols are based on an average evidence, posing challenges for selecting the best regime for an individual patient due to the diversity across or within the population. This limitation calls for RL which represent a perfect support for the discovery and generation of optimal DTRs.

Yu *et al.* (2019b) provides a detailed review on clinical and healthcare studies which used RL, including DTR estimation for specific chronic diseases. Yu *et al.* (2019b) provides a detailed review on studies which used RL in healthcare, including

DTR estimation for specific chronic diseases. However, they extensively include the multitude of works which have been developed in the DTR field, regardless their theoretical or applied nature. Notably, among these works (see Appendix B for a partial view in cancer), very few studies evaluated the proposed method in a real cohort of patients (rather than simulations). For cancer therapies, only one work exists (Tseng *et al.*, 2017). It proposes a DRL framework for estimating the optimal radiation dose escalation, considering a retrospective population of 114 non-small-cell lung carcinoma patients, and looking at different radiotherapy outcomes (reward variables).

An increased number of real-world studies can be found in chronic diseases different from cancer, such as diabetes (Yasini *et al.*, 2009; Asoh *et al.*, 2013; Luckett *et al.*, 2020), anemia (Martín-Guerrero *et al.*, 2009; Malof & Gaweda, 2011; Escandell-Montero *et al.*, 2011), HIV (Parbhoo *et al.*, 2017), substance addiction (Murphy *et al.*, 2016; Chakraborty *et al.*, 2008, 2010; Tao *et al.*, 2018), and mental health. The latter includes the popular *Sequenced Treatment Alternatives to Relieve Depression* (STAR*D) trial (Rush *et al.*, 2004) on depression, and the *Clinical Antipsychotic Trials of Intervention Effectiveness* (CATIE) study (Keefe *et al.*, 2007) on schizophrenia; based on these several RL improvements has been proposed within ML and statistics (Pineau *et al.*, 2007; Chakraborty *et al.*, 2013; Laber *et al.*, 2014b; Linn *et al.*, 2017; Schulte *et al.*, 2014; Song *et al.*, 2015; Liu *et al.*, 2018; Shortreed *et al.*, 2011; Ertefaie *et al.*, 2016; Lizotte *et al.*, 2012; Lizotte & Laber, 2016; Laber *et al.*, 2014c; Butler *et al.*, 2018). However, excluding some exceptions (Yasini *et al.*, 2009; Asoh *et al.*, 2013; Luckett *et al.*, 2020; Tao *et al.*, 2018), these studies used real data, as well as simulated data, only for evaluating the proposed RL method, thus, only as an illustrative example. In between pure simulations and real data, there's also an intermediate line of research for DTRs estimation which used either real-data to build a simulator environment (Daskalaki *et al.*, 2013; Ernst *et al.*, 2006), or established mathematical models for simulating disease specific dynamic system in patients (Ngo *et al.*, 2018), e.g., the Palumbo mathematical model (Palumbo *et al.*, 2007). We do not cover these studies.

Moving now from chronic diseases to critical care, where patients need urgent medical treatment, development of new data generating tools available for use in intensive care unit (ICU), suggests a great opportunity for applications of ML and RL methodologies (Vellido *et al.*, 2018). To date, RL has been used in the treatment of sepsis (Vellido *et al.*, 2018; Raghu *et al.*, 2017b,a; Yu *et al.*, 2019a), regulation of sedation (Moore *et al.*, 2014), and other intensive care unit problems (Wang *et al.*, 2018; Liu *et al.*, 2017). Most of these applications (a summary is provided in Yu *et al.*, 2019b) are based on the *Multiparameter Intelligent Monitoring in Intensive Care* (MIMIC)-III (Johnson *et al.*, 2016b) freely accessible database, and mainly use a DL framework for approximating the Q-learning functions. A key motivation for using the DL, is related to its higher flexibility and adaptability to high dimensional action and state spaces compared to standard RL methods and its superior capability in modelling real-life complexity in heterogeneous disease progression and treatment choices, and automatic feature extraction directly from the input data. As in the previous case, data are generally used for illustrative purposes.

The content above highlights and summarizes an increasing progress and interest in applying RL for optimal DTRs estimation. However, despite remarkable theoreti-

cal results, only a few studies applied RL for real clinical purposes. Moreover, in these cases applications simply use RL approaches (typically Q-learning) for solving DTR problems in relatively simplified settings, thus exhibiting a number of shortcomings and practical limitations, and posing interesting technical challenges and exciting open problems. To illustrate, how can we better understand and interpret the process of an RL algorithm, which often acts in a black box expressed by, for instance, deep neural networks, is a frequently debated problem, which may prevent their application in real-world. Similarly, a major concern is how to adequately adapt the RL strategy to the complex disease scenario a scientist may work in. For instance, formalizing suitable relationships in the RL process, particularly for the reward function, taking into account prior knowledge on the specific disease, multiple objectives, and presence of unstructured data. While, for instance, several software packages exist for implementing many of the reviewed algorithms (we report them in Appendix C), these are often suitable only under specific settings (e.g., only continuous and positive rewards), and require users' knowledge about the specific software.

3.1.2 Just-in-Time Adaptive Interventions in MHealth

Increasing technological sophistication and widespread use of smartphones and wearable devices provide a great platform to enhance healthcare delivery, and has led to the emergence of mobile health (mHealth; Istepanian *et al.*, 2007; Kumar *et al.*, 2013; Rehg *et al.*, 2017). MHealth is a rapidly expanding area which refers to the use of mobile technologies for conducting and managing health-related activities in an aim to support and improve healthcare at all levels of care, in both clinical and non-clinical populations. A key objective in mHealth is to deliver efficacious interventions in response to rapid changes in an individual's circumstances, while avoiding over-treatment as it leads to user *disengagement* (e.g., recommendations are ignored, app is deleted). As introduced in Section 3.1, this can be efficiently achieved by real-time AIs, known as just-in-time adaptive interventions (JITAIs; Nahum-Shani *et al.*, 2018). The distinctive feature of JITAIs, compared to DTRs, is to adapt intervention to the user's in-the-moment context or needs, e.g., time, location or current activity. This peculiarity contributed to their increasing popularity in a variety of domains, including physical activity (Consolvo *et al.*, 2008; Van Dantzig *et al.*, 2013; Hardeman *et al.*, 2019), illness management support (Ben-Zeev *et al.*, 2014), addictive disorders in alcohol and drug use (Gustafson *et al.*, 2014; Goldstein *et al.*, 2017; Garnett *et al.*, 2019; Bell *et al.*, 2020), smoking cessation (Naughton, 2017), obesity/weight management (Patrick *et al.*, 2009; Aswani *et al.*, 2019). As JITAIs are carried out in dynamic environments where context and options can change rapidly, thus requiring a continuous learning, often in indefinite horizons. Existing theory and guidance on constructing high-quality evidence-based JITAIs is still insufficiently mature to precisely specify which particular intervention and when it should be delivered. In addition, they pose some unique challenges, which we will discuss in Chapter 6 with reference to a real-world study we conduced, that preclude direct application of existing methodologies for DTRs (Nahum-Shani *et al.*, 2015, 2018).

The current standard approach for developing JITAIs is given by contextual

MABs (Tewari & Murphy, 2017), which occupy a middle ground between MAB (Bubeck & Cesa-Bianchi, 2012; Auer *et al.*, 2002b) and full-RL, as illustrated in Section 2. This is because MABs assumptions match the conceptual design of many JITAIs. In fact, intervention options in a JITAI are sometimes referred to as *Ecological Momentary Interventions* or *micro-interventions*. Such a terminology emphasizes that the effects of many of the treatments in this domain are expected to be short-lived in nature.

With a few exceptions, contextual MAB algorithms applied in mHealth rely on two fundamental MAB approaches introduced in advertising: the Linear Upper Confidence Bound (LinUCB; Li *et al.*, 2010; Chu *et al.*, 2011) and the Linear Thompson Sampling (LinTS; Agrawal & Goyal, 2013). Exceptions include the Actor-Critic strategy (Lei *et al.*, 2017) and other more full-RL oriented techniques (Zhou *et al.*, 2018).

Contextual MABs with LinUCB-based Exploration

LinUCB (Li *et al.*, 2010; Chu *et al.*, 2011), inspired by the work of Walsh *et al.* (2012), represents an extension of the upper confidence bound (UCB; Auer *et al.*, 2002a) method, in the sense that the expected reward is assumed to be a linear function of a context-action feature, say $f(X_t, A_t) \in \mathbb{R}^d$. We consider features (constructed e.g., via linear basis, polynomials or splines expansion; Marsh & Cormier, 2001) rather than a standard linear function as they may capture non-linearities in the data, yielding more predictive and explanatory power. The idea behind UCB and LinUCB is to perform 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 UCB of the expected reward value for each arm, because the interest is in the arm with the highest reward. More specifically, under the linearity assumption, i.e., $\mathbb{E}[Y_{t+1}|X_t, A_t] = f(X_t, A_t)^T\mu$, with $\mu \in \mathbb{R}^d$ the unknown coefficients vector, the proposal is to estimate the UCB associated with arm a_t at time t, say $U_t(a_t)$, by

$$\hat{U}_t(a_t) \doteq f(X_t, A_t = a_t)^T\hat{\mu}_t + \alpha s_t(a_t),$$

where the first part $f(X_t, A_t = a_t)^T\hat{\mu}_t$, with $\hat{\mu}_t \doteq B_t^{-1}b_t$ an estimator of μ, reflects the current point estimate of the reward, while the second part $s_t(a_t) \doteq \sqrt{f(X_t, A_t = a_t)^T B_t^{-1} f(X_t, A_t = a_t)}$ represents an indication of its uncertainty, i.e., the standard deviation. B_t^{-1} and b_t are analogous to the terms "$(X^TX)^{-1}$" and "X^TY", respectively, that constitute the OLS estimator in a standard linear regression model ($\mathbb{E}[Y|X] = X^T\mu$). But, here, they are recursively computed taking into account previously explored arms. Assuming a ridge penalized estimation, with penalty parameter $\lambda \geq 0$, at time t, $B_t \doteq \lambda\mathbb{I}_d + \sum_{\tau=0}^{t-1} f(X_\tau, A_\tau = \tilde{a}_\tau)^T f(X_\tau, A_\tau = \tilde{a}_\tau)$ and $b_t \doteq \sum_{\tau=0}^{t-1} f(X_\tau, A_\tau = \tilde{a}_\tau)^T Y(X_\tau, A_\tau = \tilde{a}_\tau)$, with $\{\tilde{a}_\tau \doteq \arg\max_{a_\tau \in \mathcal{A}} U_\tau(a_\tau)\}_{\tau=0,1,\ldots,t-1}$ being the estimated optimal arms on previous rounds. Algorithm 3 provides a schematic of this approach. The tuning parameter $\alpha > 0$ can be viewed as a generalization of the critical values typically used confidence intervals. It controls the trade-off between exploration and exploitation: small values of α favor exploitation while larger values of α favor exploration.

Algorithm 3: LinUCB (Li *et al.*, 2010; Chu *et al.*, 2011)

Input: $\alpha \in \mathbb{R}_+$, $\lambda \in \mathbb{R}_+$, $T \in \mathbb{N}$, $d \in \mathbb{N}$
Initialization: $B_0 = \lambda \mathbb{I}_{d'}, b_0 = \mathbf{0}_d$
for $t = 0, 1, 2, \ldots T$ **do**
  Estimate the regression coefficient $\hat{\mu}_t = B_t^{-1} b_t$;
  for $a_t \in \mathcal{A}$ **do**
    Observe feature $f(X_t, A_t = a_t)$ and compute the upper confidence bound $U_t(a_t)$

$$U_t(a_t) = f(X_t, A_t = a_t)^T \hat{\mu}_t + \alpha \sqrt{f(X_t, A_t = a_t)^T B_t^{-1} f(X_t, A_t = a_t)};$$

  end for
  Select arm $\tilde{a}_t = \arg\max_{a_t \in \mathcal{A}} U_t(a_t)$ and get the associated reward $Y_{t+1}(X_t, A_t = \tilde{a}_t)$;
  Update B_t and b_t according to the best arm $\tilde{a}_t$

$$B_{t+1} = B_t + f(X_t, A_t = \tilde{a}_t)^T f(X_t, A_t = \tilde{a}_t);$$
$$b_{t+1} = b_t + f(X_t, A_t = \tilde{a}_t)^T Y_{t+1}(X_t, A_t = \tilde{a}_t).$$

end for

Theoretical studies on LinUCB showed that they provide high probability guarantees on the regret suffered by the learner. Under the assumption on bounded features and rewards, and sub-Gaussian (Rigollet & Hütter, 2015) regression errors, Abbasi-Yadkori *et al.* (2011) showed that appropriate choices of α will give, at time T, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ a regret bound of

$$O\left(d\sqrt{T \log(T/\delta) \log(1 + T/\delta)}\right). \tag{3.16}$$

Several theoretical variations and extensions of the LinUCB have been proposed in the bandit literature. These include: i) Linear Associative RL (LinREL) and SupLinREL (Auer, 2002), based on singular-value decomposition rather than ridge regression for obtaining an estimate of the UCB; ii) generalized linear model versions (UCB-GLM; Filippi *et al.*, 2010) and SupUCB-GLM (Li *et al.*, 2017), which assumes that the reward function can be written as a composition of a linear function and a link function; iii) non-parametric modeling of the reward function, such as Gaussian processes (GP-UCB; Srinivas *et al.*, 2009, 2012); contextual GP-UCB (Krause & Ong, 2011) and kernel functions (SupKernelUCB; Valko *et al.*, 2013); iv) NeuralUCB, which leverages the representation power of DNNs and uses a neural network-based random feature mapping to construct the UCB for the reward (Zhou *et al.*, 2019).

More recently, in addition to the exploration-exploitation dilemma, other statistical objectives started to be considered in the LinUCB. To accommodate more complex models of the world, Urteaga & Wiggins (2018), propose leveraging advances in sequential Monte Carlo (SMC) methods from the approximate inference community. More specifically, they incorporate the flexibility of (sequential) importance sampling to allow for accurate estimation of statistics of interest which cannot be computed in closed-form, within the MAB problem. By combining SMC methods -

which estimate posterior densities and expectations in probabilistic models that are analytically intractable - with Bayesian UCB-based algorithms, they extend their applicability to complex models: those for which sampling may be performed even if analytic computation of summary statistics is infeasible, e.g., non linear reward functions and dynamic bandits. In a similar context to ours, i.e., behavioural science, Dimakopoulou *et al.* (2019) introduced balancing methods from the causal inference literature, i.e., weighting each observation with the estimated inverse probability of a context being observed for an arm, in the regression estimation process, in order to make the bandit algorithm less prone to bias. More specifically, the idea is that at every time t, the linear contextual bandit weighs each observation $(x_\tau, a_\tau, y_{\tau+1})$, $\tau = 1, \ldots, t$ in the history up to time t by $\hat{w}_{a_\tau} = 1/\hat{p}_{a_\tau}(x_\tau)$, i.e., the estimated inverse probability of context x_τ being assigned to arm a_τ. Theoretical guarantees of their *Balanced UCB* (BLUCB) match the state-of-the-art guarantees, but it helps to reduce bias, particularly in misspecified cases, at a cost of increased variance. Clipping the propensity scores away from zero (Crump *et al.*, 2009) with some threshold, e.g. 0.1 can control the variance increase.

Real-world studies. Moving now from to real-world mHealth applications, the use of LinUCB has been encountered in Paredes *et al.* (2014) and Forman *et al.* (2019). The former developed a LinUCB based intervention recommender system for learning how to match interventions to individuals and their temporal circumstances over time. The aim was to send one of the 10 available types of stress management strategies (upon user's request in the mobile app), with the goal of maximizing stress reduction. After four weeks of study, participants receiving the LinUCB-based recommendations showed a tendency towards using more constructive coping behaviors. Similarly, Forman *et al.* (2019), in the context of behavioural weight loss (WL) and maintenance, conducted a pilot experimental study to evaluate the feasibility and acceptability of an RL-based WL intervention system, and whether it would achieve equivalent benefit at a reduced cost, compared to a non-optimized intervention system. To this purpose, participants were randomized between a non-optimized, a individually optimized (individual reward maximization) and a group optimized (group reward maximization) group. Reward optimization was based on the UCB technique Auer *et al.* (2002a) and adjusted for intervention cost (i.e., time): based on previous work (Ontanón, 2013, 2017), UCB was used to balance the need for exploiting the best intervention (thus far) with the need for exploring interventions that had not been explored for a long time for a particular person. Specifically, the likelihood that a lower-reward-score intervention would be explored for a participant was proportional to how many days it had been since the last time this intervention had been delivered to that particular person. The study showed that the LinUCB-based optimized groups have strong promise in terms of outcome of interest, not only being feasible to deploy and acceptable to participants and coaches, but also achieved desirable results at roughly one-third the cost.

Contextual Bandits with LinTS-based Exploration

Under the same linear reward function assumption considered above, Agrawal & Goyal (2013) proposed a randomized version of LinUCB, based on a generalization

of the Thompson Sampling (TS) technique for stochastic contextual MABs problems. Rooted in a Bayesian framework, the idea of TS is to select an arm according to its posterior probability of being optimal, i.e., maximizing the posterior reward distribution. More specifically, assuming a Gaussian prior for the μ regression coefficients vector, e.g., $\mu \sim \mathcal{N}(\mathbf{0}_d, \sigma_\mu^2 \mathbb{I}_d)$, and a Gaussian distribution for the reward, i.e., $Y_t|\mu, f(X_t, A_t) \sim \mathcal{N}(f(X_t, A_t)^T\mu, \nu^2)$, at each time t, the optimal arm $\tilde{a}_t$ will be the one that maximises the a-posteriori estimated expected reward, i.e., $f(X_t, A_t)^T\tilde{\mu}_t$. The posterior nature is reflected in $\tilde{\mu}_t$, which represents a sample from the posterior distribution, given by $\mathcal{N}(\hat{\mu}_t, \nu^2 B_t^{-1})$; here $\hat{\mu}_t \doteq B_t^{-1}b_t$ is the posterior mean, with B_t and b_t defined in the same way as for LinUCB. The full iterative process is described in Algorithm 4.

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

Authors showed that TS achieves strong regret guarantees: with probability $1-\delta$, $\delta \in (0,1)$, the total regret for TS at time T is bounded as

$$O\left(\frac{d^2}{\epsilon}\sqrt{T^{1+\epsilon}}(\log(Td)\log(\frac{1}{\delta}))\right), \tag{3.17}$$

for any $\epsilon \in (0,1)$ tuning parameter of the algorithm. If T is known, Agrawal & Goyal (2013) suggest to choose $\epsilon = \frac{1}{\log T}$ to get $\tilde{O}(d^2\sqrt{T})$ regret bound.

Algorithm 4: LinTS (Agrawal & Goyal, 2013)

Input: $\sigma_\mu \in \mathbb{R}$, $\nu \in \mathbb{R}$, $T \in \mathbb{N}$, $d \in \mathbb{N}$, $\lambda \in \mathbb{R}_+$
Initialization: $B_0 = \lambda \mathbb{I}_d, b_0 = \mathbf{0}_d$
for $t = 0, 1, 2, \dots T$ **do**
  Estimate the regression coefficient $\hat{\mu}_t = B_t^{-1}b_t$;
  Get posterior samples $\tilde{\mu}_t \sim \mathcal{N}(\hat{\mu}_t, \nu^2 B_t^{-1})$
  for $a_t \in \mathcal{A}$ **do**
    Observe feature $f(X_t, A_t = a_t)$ and compute the 'a-posteriori' estimated expected reward, i.e., $f(X_t, A_t = a_t)\tilde{\mu}_t$
  end for
  Select arm $\tilde{a}_t = \arg\max_{a_t \in \mathcal{A}} f(X_t, A_t = a_t)\tilde{\mu}_t$ and get the associated reward $Y_{t+1}(X_t, A_t = \tilde{a}_t)$;
  Update B_{t+1} and b_{t+1} according to the best arm $\tilde{a}_t$

$$B_{t+1} = B_t + f(X_t, A_t = \tilde{a}_t)^T f(X_t, A_t = \tilde{a}_t);$$
$$b_{t+1} = b_t + f(X_t, A_t = \tilde{a}_t)^T Y_{t+1}(X_t, A_t = \tilde{a}_t).$$

end for

Similarly to LinUCB, LinTS has been intensively studied within the theoretical literature, and several extensions, including those proposed by Dimakopoulou *et al.* (2019) and Urteaga & Wiggins (2018) for LinUCB, and a generalized TS version (Li, 2013), based on more general loss functions, have been considered. The latter introduced a new family of algorithms, called *Generalized Thompson Sampling for Contextual Bandits*, and analyze its regret in the expert-learning framework, where each expert corresponds to a contextual policy for arm selection. Motivated by the TS's Bayes update rule, which can be viewed as an exponentiated update with the logarithmic loss, the idea is to use a more general loss function to update the experts' weights by comparing the actual received reward and the expert prediction. Then, similarly to the TS idea, the Generalized TS follows an expert's policy more often if the expert is more likely to be optimal.

In what follows we focus on works which have been developed within the mHealth literature, specifically addressing field-related characteristics.

Bootstrap Thompson Sampling. A different extension, more targeted to efficient computation of the posterior distribution of the TS, is proposed in Eckles & Kaptein (2014) and Eckles & Kaptein (2019). Under the normal conjugate family in Algorithm 4, sampling from the posterior is straightforward. However, to be practically feasible for many problems, and thus scalable to large T or to complex likelihood functions, TS requires computationally efficient sampling from the distribution of the regression parameters $\mu_t|\mathcal{T}_{t-1}, f(X_t, A_t)$. Already in situations in which a logit or probit model is used to model the expected reward of the actions, the posterior is not available in closed form and is then often computed using Markov chain Monte Carlo (MCMC) methods or otherwise approximated, which can be computationally costly. To this end, motivated by relationships between bootstrap distributions Rubin (1981) and Bayesian posteriors, and by the fact that the bootstrap distribution can be used to approximate posteriors Efron (2012); Newton & Raftery (1994), in Eckles & Kaptein (2019), a *Bootstrap Thompson Sampling* (BTS) technique for replacing the posterior by an online bootstrap distribution of the point estimate $\hat{\mu}_t$ at each time t, is proposed. Specifically, in BTS, at each step t, the algorithm first chooses the best arm according to a single, uniformly randomly selected, bootstrap replicate, and then, for each bootstrap replicate $j = 1, \ldots, J$, each observation i gets an updating weight $w_{ij} \sim 2 \times \text{Bern}(1/2)$ (see McCarthy (1969); Owen *et al.* (2012) for more details on reweighting bootstrap). In an empirical evaluations, authors showed that, in comparison with LinTS and other methods, BTS it is more robust to some kinds of model misspecifications, thanks to the robustness of the bootstrap for statistical inference, and it can be easily adapted to dependent observations, a common feature of behavioral sciences (Eckles & Kaptein, 2014, 2019).

Intelligent-Pooling Thompson Sampling. When data on individuals is limited, learning an adaptive strategy separately for each user may be a very slow process, particularly if data are sparse and/or noisy and the process is non-stationary. However, a natural tension exists between individual personalizing and pooling using data over users, a choice which can introduce bias. For balancing this tension, Tomkins *et al.* (2019) introduced a novel algorithm called *Intelligent Pooling* that generalizes LinTS by using a Gaussian mixed effects linear model for the reward. Mixed effects models are widely used across behavioral sciences, including mHealth

(Qian *et al.*, 2020), to model the variation in the linear model parameters across individual and within an individual across time (Raudenbush & Bryk, 2002; Laird & Ware, 1982). The idea of this method is to use random effects to adaptively pool users' data based on the degree to which users exhibit heterogeneous rewards. In addition, unlike LinTS, in which the prior hyper-parameters are set at the beginning, their proposal includes a procedure for updating the hyper-parameters online. Empirical evaluations showed that Intelligent Pooling achieves an average of 26% lower regret than state-of-the-art, demonstrating promise of personalization on even a small group of users.

Action-Centered Thompson Sampling. Moving now to specific challenges arising in emerging mHealth applications, Greenewald *et al.* (2017) provides an extension of the linear model for contextual bandits by specifically targeting the time-invariant and linear model assumption, which is untenable in human behavior. They consider a particular type of non-stationary and non-linear contextual bandit that has two parts: a baseline reward (associated with a "do nothing" or control arm, denoted with 0) and a treatment or action effect. Assuming K (non-control) arms, in addition to the 0 (control) arm, at each time step $t \in \mathbb{N}$, the expected reward model is formalized as

$$\mathbb{E}(Y_{t+1}|X_t = x_t, A_t = a_t) = f(x_t, a_t)^T \mu \mathbb{I}(a_t \neq 0) + g_t(x_t), \tag{3.18}$$

with $f(x_t, a_t) \in \mathbb{R}^d$ a fixed context-action feature (with context X_t chosen by an adversary on the basis of the trajectory $\mathcal{T}_{t-1}$ up to time t), $\mu \in \mathbb{R}^d$ the parameters vector, and $g_t(x_t)$ a time-varying component that can vary in a way that depends on the past, but does not depend on current action (thus, allowing for non-stationarity). The term adversarial in contextual bandits refers to the context and reward generation mechanism: when both contexts and actions are allowed to be chosen arbitrarily by an adversary, no assumptions on generating process nature are made (Tewari & Murphy, 2017). Note that, due to the indicator function $\mathbb{I}(a_t > 0)$, the expected reward when baseline action 0 is taken is given by the time-varying component only, which we regard as a baseline reward. Based on this framework, a linear model for the reward difference, or differential reward, at time t is obtained as

$$Y_{t+1}(X_t, A_t) - Y_{t+1}(X_t, 0) = f(X_t, A_t)^T \mu \mathbb{I}(A_t > 0) + \epsilon_{t+1},$$

where ϵ_{t+1} is zero-mean sub-Gaussian noise with variance σ^2. If the bandit had access at each time t to the differential reward, estimating the unknown parameter μ would be straightforward, e.g., by using the ordinary or penalized least-squares approach, as seen in LinUCB and LinTS. However, we only have access to the observed $Y_{t+1}(X_t, A_t)$, which contains the sum of the arbitrarily complex baseline reward and the differential reward we want to estimate. To isolate the differential reward at each time step, authors propose the so called *action-centering trick*, which randomizes the action at each time step, allowing to construct an estimator whose expectation is proportional to the differential reward $Y_{t+1}(X_t, \overline{A}_t) - Y_{t+1}(X_t, 0)$, where $\overline{A}_t$ is the nonzero action chosen by the bandit at time t to be randomized against the zero action. More specifically, denoted with $\pi_t \doteq 1 - \pi_t(0, t) \doteq \mathbb{P}(A_t > 0)$

the probability of taking a non-zero action, we have that

$$\begin{aligned}\Big[(\mathbb{I}(A_t > 0) - \pi_t)Y_{t+1}(X_t, A_t)|\mathbf{H}_t, A_t\Big] &= \pi_t(1-\pi_t)Y_{t+1}(X_t, \overline{A}_t) \\ &\quad - (1-\pi_t)\pi_t Y_{t+1}(X_t, 0) \\ &= \pi_t(1-\pi_t)\big(Y_{t+1}(X_t, \overline{A}_t) - Y_{t+1}(X_t, 0)\big),\end{aligned}$$

meaning that the observed reward $Y_{t+1}(X_t, A_t)$ is proportional to an unbiased estimator of $Y_{t+1}(X_t, \overline{A}_t) - Y_{t+1}(X_t, 0)$. Thus, the proposed estimate of the differential reward at time t, which we call pseudo-reward, is given by $\hat{Y}_{t+1}(X_t, \overline{A}_t) = \mathbb{I}(A_t > 0) - \pi_t)Y_{t+1}(X_t, A_t)$. An important property of this pseudo-reward is that its conditional expectation does not depend on the arbitrarly complex $g_t(x_t)$ term.

Furthermore, to avoid sending too few or too many interventions, and prevent the algorithm from converging to an ineffective deterministic policy, a constraint on the size of the probabilities of delivering a non-control intervention (i.e., probability clipping) is considered:

$$0 < \pi_{min} \leq 1 - \pi(A_t = 0|X_t) \leq \pi_{max} < 1,$$

where $1 - \pi(A_t = 0|X_t)$ is the conditional bandit-chosen probability of delivering an intervention at time t, and the constants π_{min} and π_{max} in $[0, 1]$ are not learned by the algorithm, but chosen using domain science, and might vary for different components of the same mHealth system.

In the context of a LinTS strategy, the proposed *Action-Centered Thompson Sampling* (ACTS) method can be viewed as a two-step hierarchical procedure, where the first step, is to estimate the arm that maximizes the reward, and the second step is to randomly determine whether to take the non-control arm, choosing an arm $A_t \neq 0$ with probability π_t given by

$$\pi_t = \mathbb{P}(A_t \neq 0) = \max\Big(\pi_{min}, \min(\pi_{max}, \mathbb{P}(f(X_t, \overline{A}_t)^T\tilde{\mu} > 0)\Big), \tag{3.19}$$

where $\overline{A}_t$ denotes a random non-control arm, $\tilde{\mu}$ a draw from the posterior distribution defined in the TS algorithm, and π_{min} and π_{max} in $[0, 1]$ are constant probability constraints. This procedure is summarized in Algorithm 5.

Under restrictive conditions on the action choice probabilities, authors showed that ACTS achieves performance guarantees similar to the linear reward setting, while still allowing for non-linearities in the baseline reward. Empirical evaluations on a popular mHealth study, known as *HeartSteps*, were also performed. HeartSteps is an app and MRT (Liao *et al.*, 2015; Klasnja *et al.*, 2015) aiming to evaluate the efficacy of contextually tailored activity suggestions and activity planning for increasing physical activity among sedentary adults (Klasnja *et al.*, 2015, 2019). It has been the subject of interest of many works in both biostatistics (Liao *et al.*, 2016; Boruvka *et al.*, 2018) and RL/bandit literature (Greenewald *et al.*, 2017; Lei *et al.*, 2017; Liao *et al.*, 2020). Following the ACTS strategy, Liao *et al.* (2020), for instance, incorporated in the differential reward model an "availability" variable, stating whether the user is available to receive an intervention.

Further theoretical improvements over the ACTS can be found in both Krishnamurthy *et al.* (2018) and Kim & Paik (2019). Here, a relaxation of the action-independent assumption of the component $g_t(x_t)$ in (3.1.2) of the ACTS is considered,

Algorithm 5: Action-Centered TS (Greenewald *et al.*, 2017)

Input: $\nu \doteq R\sqrt{9d' \log(T/\delta)}$, $T \in \mathbb{N}$, $d' \in \mathbb{N}$, $(\pi_{min}, \pi_{max}) \in [0,1]$
Initialization: $B_0 = \mathbb{I}_{d'}, b_0 = \mathbf{0}_{d'}$
for $t = 0, 1, 2, \dots T$ **do**
  Estimate the regression coefficient $\hat{\mu}_t = B_t^{-1} b_t$;
  Get posterior samples $\tilde{\mu}_t \sim \mathcal{N}(\hat{\mu}_t, \nu^2 B_t^{-1})$
  for $\overline{a}_t = \mathcal{A} \doteq \{1, \dots, K\}$ **do**
    Observe feature $f(X_t, \overline{A}_t = \overline{a}_t)$ and compute the 'a-posteriori' estimated expected reward, i.e., $f(X_t, \overline{A}_t = \overline{a}_t)\tilde{\mu}_t$
  end for
  Let $\overline{a}_t^* = \arg\max_{\overline{a}_t \in \mathcal{A}} f(X_t, \overline{A}_t = \overline{a}_t)\tilde{\mu}_t$;
  Compute the probability π_t of taking non-zero action and play action $\tilde{a}_t = \overline{a}_t^*$ with probability π_t, else play $\tilde{a}_t = 0$;
  Get the associated reward $Y_{t+1}(X_t, \overline{A}_t = \tilde{a}_t)$;
  Update B_{t+1} and b_{t+1} according to arms $\overline{a}_t^*$ and $\tilde{a}_t$

$$B_{t+1} = B_t + \pi_t(1-\pi_t) f(X_t, A_t = \overline{a}_t^*) f(X_t, A_t = \overline{a}_t^*)^T;$$
$$b_{t+1} = b_t + (\mathbb{I}(A_t > 0) - \pi_t) f(X_t, A_t = \overline{a}_t^*) Y_{t+1}(X_t, A_t = \tilde{a}_t).$$

end for

making the reward model at time t entirely non-parametric when allowing dependence on both time and history, i.e., $\mathbb{E}(Y_{t+1}|\mathbf{H_t} = \mathbf{h_t}, A_t = a_t) = f(x_t, a_t)^T \mu + g_t$. For estimating the unknown parameters, Krishnamurthy *et al.* (2018) proposed the adversarial *Bandit Orthogonalized Semiparametric Estimation* (BOSE) method, based on an action-elimination strategy adapted from Even-Dar *et al.* (2006), and a centering trick as in Greenewald *et al.* (2017) to cancel out g_t. The proposed estimator $\hat{\mu}_t$ at time t is given by

$$\hat{\mu}_t = \left(\lambda \mathbb{I}_d + \sum_{\tau=0}^{t-1} \overline{f}(X_\tau, A_\tau) \overline{f}(X_\tau, A_\tau)^T\right)^{-1} \sum_{\tau=0}^{t-1} \overline{f}(X_\tau, A_\tau) Y_{\tau+1}(X_\tau, A_\tau),$$

where $\lambda \geq 0$ is the ridge penalty parameter, and $\overline{f}(X_\tau, A_\tau) \doteq f(X_\tau, A_\tau) - \mathbb{E}(f(X_\tau, A_\tau)|\mathbf{H}_\tau, A_\tau)$ represents the centering trick. It is derived so that, conditionally on $(\mathbf{H}_\tau, A_\tau)$, $\mathbb{E}(\overline{f}(X_\tau, A_\tau)|\mathbf{H}_\tau, A_\tau) = \mathbf{0}_d$.

The BOSE algorithm does not require any constraint on the action choice probabilities as in (3.19), and it achieves a $(1-\delta)$-probability regret bound of $O(d\sqrt{Tlog(T/\delta)})$, that matches the best known regret bound of LinUCB for linear reward models (see 3.16). However, the action elimination step requires $O(K^2)$ computations at each round, and, in order to meet the regret bound, the distribution used to select the action should satisfy specific non-trivial conditions. To overcomes this difficulties, under the same framework, Kim & Paik (2019) proposed

an alternative estimator, given by

$$\hat{\mu}_t = \Big(\mathbb{I}_d + \sum_{\tau=0}^{t-1} \overline{f}(X_\tau, A_\tau)\overline{f}(X_\tau, A_\tau)^T + \sum_{\tau=0}^{t-1} \mathbb{E}(\overline{f}(X_\tau, A_\tau)\overline{f}(X_\tau, A_\tau)^T | \mathbf{H}_\tau, A_\tau)\Big)^{-1} \sum_{\tau=0}^{t-1} 2\overline{f}(X_\tau, A_\tau) Y_{\tau+1}(X_\tau, A_\tau).$$

The proposed algorithm requires only $O(K)$ computations at each round, and enjoys a tighter high-probability upper bound than the BOSE. This bound matches the bound of the LinTS algorithm reported in (3.17).

Actor-Critic Contextual Bandits

Specifically addressing the problem of personalized mHealth interventions, Lei (2016) and Lei *et al.* (2017) proposed to use an alternative class of RL algorithms, known as *actor-critic* (AC) RL (Sutton & Barto, 2018; Grondman *et al.*, 2012), based on which both policies and value functions are directly learned. *Actor* is the component that learns policies, and *critic* the one that learns a value function, which is then used to "critizise" and update actor's policy in a direction of performance improvement. In this sense, AC architectures combine direct and indirect methods, and in specific settings they provide a framework of equivalence for the two distinct approaches (Guan *et al.*, 2019).

Considering a binary action space $\mathcal{A} = \{0, 1\}$, and assumptions similar to the ones of LinTS and LinUCB (i.e., linear reward model and bounded rewards and features), authors formulated the problem as a stochastic contextual MAB and proposed a class of parameterized stochastic policies, with $\mathbb{P}(A = 1|X = x) = \pi(1|x;\theta) = \frac{e^{g(x)^T\theta}}{1+e^{g(x)^T\theta}}$, and $g(x)$ a p-dimensional policy feature. Similarly to the ACTS algorithm (Greenewald *et al.*, 2017) illustated in Section 3.1.2, authors also considered a stochastic chance constraint of the form

$$\mathbb{P}(\pi_{\min} \leq \pi(A = 1|X;\theta) \leq 1 - \pi_{\min}) \geq 1 - \alpha, \tag{3.20}$$

with $\pi_{\min} \in (0, .5)$ and $\alpha \in (0, 1)$ being constants controlling the amount of stochasticity. By improving treatment variety, this constraint may increase engagement and decrease the *habituation* effect (Raynor & Epstein, 2001; Epstein *et al.*, 2009; Wilson *et al.*, 2005) which can incur in deterministic policies.

An optimal policy is then obtained by maximizing the expected reward under the policy $\pi(a|x;\theta)$, i.e., $V(\theta) \doteq \mathbb{E}_{\pi_\theta}(Y)$, subject to the constraint in (3.20). Solving this constrained optimization problem involves a major difficulty since it is, in general, a non-convex constraint on θ, involving also some non-smoothness. To circumvent this difficulty, first, a relaxation of (3.20) is made, and then the Lagrangian function $J_\lambda(\theta)$, with λ the Lagrangian multiplier, is proposed as an alternative objective, referred to as *regularized average reward*. That is,

$$\begin{aligned} J_\lambda(\theta) &\doteq V(\theta) - \lambda\theta^T\mathbb{E}(g(X)g(X)^T)\theta \\ &= \mathbb{E}_{\pi_\theta}(Y) - \lambda\theta^T\mathbb{E}(g(X)g(X)^T)\theta \\ &= \mathbb{E}_{p(x)}\mathbb{E}_{\pi(a|x;\theta)}[E(Y|X = x; A = a)] - \lambda\theta^T\mathbb{E}(g(X)g(X)^T)\theta, \end{aligned} \tag{3.21}$$

where $p(x)$ is a fixed unknown distribution of the context. For a given λ, the optimal policy $\pi^* \doteq \pi_{\theta^*}$ is the one with $\theta^* \doteq \arg\max_{\theta\in\Theta} J_\lambda(\theta)$. However, both $J_\lambda(\theta)$ and $E(Y|X=x;A=a)$ are unknown. The conditional mean of the reward or Q-function is first estimated through a penalized (L_2-norm) linear regression as in LinTS and LinUCB - this is the critic step. Then, the obtained estimates for each a and x are plugged into (3.21) and an estimated optimal actor's policy is derived based on the Monte-Carlo estimator:

$$\widehat{J}_\lambda(\theta) \doteq \mathbb{P}_T \left[\sum_a \hat{\mathbb{E}}(Y|X=x;A=a)\pi(a|X=x;\theta) - \lambda\theta^T(g(x)g(x)^T)\theta \right], \quad (3.22)$$

where $\mathbb{P}_T$ denotes the empirical average on T i.i.d. samples. We illustrate the full procedure in Algorithm 6.

Algorithm 6: Actor-Critic Contextual Bandits (Lei *et al.*, 2017)

Input: $T \in \mathbb{N}$, $\lambda \in \mathbb{R}_+$, a class of parameterized policies $\{\pi_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ based on a p-dimensional policy feature $g(x)$

Critic Initialization: $B_0 = \lambda\mathbb{I}_{d'}, b_0 = \mathbf{0}_{d'}$

Actor Initialization: Initial policy parameter $\hat{\theta}_0$ based on domain theory or prior data

for $t = 0, 1, 2, \ldots T$ **do**

Observe context X_t and the feature vector $f(X_t, A_t)$;

Draw an action $\tilde{a}_t$ according to probability distribution $\pi_{\hat{\theta}_t}(X_t, A_t)$;

Get the associated reward $Y_{t+1}(X_t, A_t = \tilde{a}_t)$;

Critic Updates: update

$$B_{t+1} = B_t + f(X_t, A_t = \tilde{a}_t)f(X_t, A_t = \tilde{a}_t)^T;$$
$$b_{t+1} = b_t + f(X_t, A_t = \tilde{a}_t)Y_{t+1}(X_t, A_t = \tilde{a}_t);$$

and estimate the regression coefficient $\hat{\mu}_t = B_{t+1}^{-1}b_{t+1}$ and the associated reward $\hat{Y}_{t+1} = f(X_t, A_t = \tilde{a}_t)\hat{\mu}_t$

Actor Update: estimate the unknown policy parameter θ

$$\hat{\theta}_t = \arg\max_{\theta\in\Theta} \hat{J}_\lambda(\theta) = \frac{1}{t}\sum_{\tau=0}^{t}\left[\sum_a \hat{Y}_{\tau+1}\pi(a|X_\tau;\theta) - \lambda\theta^T(g(X_\tau)g(X_\tau)^T)\theta\right]$$

end for

An extension of the AC contextual bandit algorithm of Lei (2016) and Lei *et al.* (2017), was later proposed in Zhu *et al.* (2018), with the aim of including potential presence of outliers in the data trajectory of mhealth applications. Their proposal involve, first, during the critic step, the use of the *capped*-L_2 norm instead of the standard L_2 norm measure for estimating the expected rewards. That is, iteration i, the feature parameter estimates are obtained as

$$\hat{\mu}_\epsilon^{(i)} = \left(\sum_{t=0}^{T} f(X_t, A_t)U_t^{(i-1)}f(X_t, A_t)^T + \lambda\mathbb{I}_p\right)^{-1}\sum_{t=0}^{T} f(X_t, A_t)U_t^{(i-1)}Y_{t+1}, \quad (3.23)$$

where the weights $U_t^{(i-1)} \doteq \mathbb{I}(||Y_{t+1} - f(X_t, A_t)^T \hat{\mu}^{(i-1)}||_2^2 \leq \epsilon)$, for $t = 0, 1, \ldots, T$, specify which tuple should be treated as an effective observation and which as an outlier, i.e. given a fixed threshold $\epsilon > 0$, if the residual of tuple t is greater than ϵ, then we regard the t-th tuple as an outlier. Convergence of estimator in (3.23) is reached after a finite number of iterations.

Then, to boost also robustness on the actor step's objective $J_\lambda(\theta)$, the estimated weights U_t, $t = 0, 1, \ldots, T$, learned in the critic step, are considered. Recalling an estimate of $J_\lambda(\theta)$ is obtained through (3.22) by plugging in the estimated expected reward, its capped-L_2 norm robust version is given by

$$\widehat{J}_{\lambda,\epsilon}(\theta) = \frac{T_\epsilon}{T} \times \mathbb{P}_{T_\epsilon} \left[\sum_a f(X = x, A = a)^T \hat{\mu}_\epsilon \pi(a|X = x; \theta) - \lambda \theta^T (g(x) g(x)^T) \theta \right],$$

where $\mathbb{P}_{T_\epsilon}$ denotes the empirical average on the $T_\epsilon \leq T$ i.i.d. samples of X whose residuals satisfy the ϵ-capped condition, or equivalently, for which U_t is equal to 1. Compared to Lei *et al.* (2017) actor's objective in (3.22), an extra weight term U_t here is added. It gives those tuples, whose residuals are very large in the critic update, zero weight, thus are not considered in the actor updating.

Other RL-based Approaches used in MHealth Interventions

The majority of mHealth studies that used RL, or, more specifically MAB algorithms, focused on the popular Contextual MAB strategies of UCB, TS and Actor-Critic. There exist, however, other mHealth applications which used RL techniques falling in categories different from the ones just mentioned. These include the works of Yom-Tov *et al.* (2017), for evaluating the effectiveness of personalized feedback in increasing adherence of diabetic patients to recommended physical activity regimes; Zhou *et al.* (2018), for developing a fitness app, *CalFit*, which automatically sets personalized, adaptive daily step goals and adopts behavior-change features such as self-monitoring; and Rabbi *et al.* (2015), again for developing a physical activity app, *MyBehavior*, able to automatically learn users' physical activity and dietary behavior, and strategically suggest changes to those behaviors for a healthier lifestyle, also incorporating users' preferences. While the first two works are based on a more full-oriented RL, the last one considers an adversarial bandit approach, namely the randomized context-free *exponential-weight algorithm for exploration and exploitation* (EXP3; Auer *et al.*, 2002b; Bubeck & Cesa-Bianchi, 2012). In EXP3, most beneficial actions are frequently exploited with seldom exploration of less beneficial ones. It has been shown to be able to quickly adapt to changes in underlying payoff functions, meaning that, if the user starts following new suggestions or his/her lifestyle changes (e.g., moving to a new location), then underlying benefits of certain behavior also change.

In Yom-Tov *et al.* (2017) two policies were considered for the treatment arms: an "initial policy" based on the results of Elliot & Church (1997) to incentivise exploration, designed so that i) no message was sent on 20% of days, and ii) for the remaining days, a negative or a positive feedback might be received by the user with equal probability based on their expected fraction of activity; and a "learning policy", based on a linear regression algorithm with interactions and the Bolzmann

sampling (Watkins, 1989) on the outputs of the learning algorithm to choose the feedback message to be given.

Finally, Zhou *et al.* (2018) propose a predictive quantitative model for each participant based on the historical steps and goal data for that user, as in Aswani *et al.* (2019), and a two-stage RL for selecting the optimal interventions: in the first stage, inverse reinforcement learning (Ng *et al.*, 2000) is employed to estimate the parameters of the assumed predictive model for each user; in the second stage, an RL technique equivalent to a direct policy search (Sutton & Barto, 2018), using the model parameters estimated in the first stage, is used.

3.2 RL for Designing Adaptive Clinical Trials

Well-designed randomized controlled trials (RCTs) have long been recognised as the gold standard for conducting evidence-based clinical research for assessing efficacy or effectiveness of interventions. The traditional way of conducting an RCT, and more generally any clinical trial, is by following an underlying fixed design, where with fixed we mean that key elements of a design, e.g., sample size or randomization probabilities, are not allowed to change during the course of the trial. These are typically determined according to study objectives and main hypothesis (Friedman *et al.*, 2015), so that certain statistical properties are guaranteed. Once the pre-specified sample size is reached and the study ends, collected data are used for final analyses.

While these types of designs are valued for their strong statistical guarantees, they do not give the investigator the flexibility of making desirable or necessary changes based on continuously emerging knowledge of an ongoing trial (Pallmann *et al.*, 2018). For example, an interim analysis may provide enough evidence for stopping the trial earlier for success or lack of efficacy. Such decision has also a relevant ethical component, since there is a responsibility to minimise the number of people given an unsafe, ineffective, or clearly inferior treatment. In addition, traditional RCTs can demand substantial time and resources, in terms of both sample size and cost. These limitations have been widely acknowledged as limiting medical innovation (Bothwell *et al.*, 2016).

Adaptive trial designs have been proposed as a means to increase the flexibility and efficiency of RCTs, extending the benefits to trial participants, in addition to future patients, and advancing patient care by enhancing the likelihood of finding a true benefit, and reducing costs (Bhatt & Mehta, 2016). The fundamental characteristic of an adaptive clinical trial is to dynamically adjust key elements of the underlying adaptive design (e.g., randomization probabilities, sample size or compared treatments) while patients enrollment in the trial is ongoing. Adaptive designs (ADs) define the potential changes of the ongoing trial and schedule the interim looks at the data, so that its integrity and validity is not compromised (Chow *et al.*, 2005; Pallmann *et al.*, 2018). In Fig. 3.6 an illustrative comparison between traditional clinical trials and adaptive clinical trials is provided.

As pointed out by FDA (2019), each modification should be "prospectively planned". It is thus essential to distinguish prospective (i.e., by design or pre-planned) from unplanned ad hoc changes, that may commonly occur in ongoing

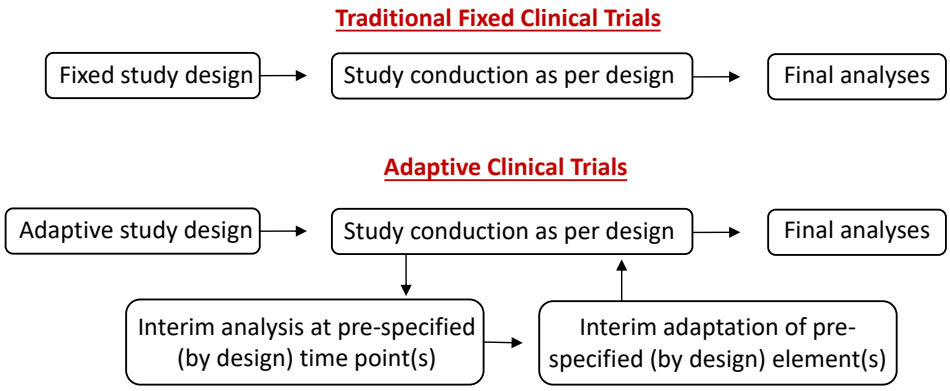

Figure 3.6. Illustrative comparison between traditional fixed clinical trials and adaptive clinical trials. An adaptive study design, as opposed to a fixed study design, allows for modifications of design elements such as sample size or randomization probabilities, based on interim data.

traditional trials (e.g., alterations to the eligibility criteria). To this purpose, a recent surge of research focuses on the quality of randomized adaptive clinical trials, arguing for an improved transparency and reporting, in order to allow results interpretability and methods reproducibility (Dimairo *et al.*, 2020), especially with the rise of new (more complex) methodological strategies.

The use of adaptive designs for modifying aspects of ongoing clinical trials has been discussed and practiced for years in clinical research. The concept can be traced back to the seminar paper of Thompson (1933). Thompson introduces an adaptive randomization method and debates the potential "considerable saving of individuals otherwise sacrificed to the inferior treatment" when a superior treatment is tried against an inferior alternative. Due to this historical link-up, the term adaptive design is sometimes used to specifically refer to adaptive randomization (Efron, 1971a; Lachin, 1988; Rosenberger *et al.*, 2001). However, the notion applies more generally to any trial design that allows some type of adaptive modification (not only the randomization probability) of an ongoing trial based on interim analyses (Bhatt & Mehta, 2016). In Table 3.3 we provide an overview of well-recognized ADs summarizing their underlying idea and the adaptive nomenclature, while a rich discussion on the topic is given in Chow & Chang (2008). Note that a single trial may also consider multiple adaptations, e.g., a group-sequential design featured with a mid-course sample size re-estimation and/or adaptive randomisation. To illustrate, consider a *seamless phase-II/III design*, representing a fusion of a phase-II (e.g., treatment selection) and phase III (e.g., efficacy confirmation) design. An adaptive seamless phase-II/III design may define three stages and two interim analyses before the end of the enrollment such that: (1) after the exploratory stage (phase II), in which subjects are assigned equally to e.g. five treatments (four active and one placebo), a first interim analysis is performed to select the best active treatment; then, (2) at a second stage (confirmatory phase III), the best active treatment and the placebo groups continue, while new patients are allocated equally to the two arms, and an interim analysis based on second-stage response-data is performed to skew the allocation probabilities in favour to the most successful treatment; finally, (3) during the third stage (confirmatory phase III), the best active treatment and

the placebo arm are allocated according to the adjusted probabilities, and second- and third-stage data are used for efficacy confirmation in the final analyses. A main motivation behind adaptive seamless phase-II/III designs is the possibility of shorten the time and patient exposure in the development of new needed drugs, by efficiently (and more ethically) use the collected data (Maca *et al.*, 2006). Carrying forward only a sub-number of the initially compared treatments may provide relevant improvement of the power of the phase-III confirmatory phase trial.

Once the idea of an adaptive design is clear to the reader, they may wonder how the adaptation is performed in order to (optimally) meet the objective of such a trial. Traditionally, rule-based or standard Bayesian and frequentist statistical approaches were considered for guiding the adaptation of adaptive clinical trial. For example, in adaptive dose-finding phase-I designs with the objective of determining the *maximum tolerated dose* (MTD), a common rule-based strategy is the so called "3-plus-3" dose-escalation design. This design is implemented as follows: three patients are initially enrolled into a given dose cohort, and, if no toxicities are observed, a dose escalation occurs for the subsequent three patients; otherwise an additional three patients are treated at the same dose level. If only one of the six enrolled patients has toxicity, escalation again continues; otherwise the trial stops, with the lower dose declared as MTD.

More recently, machine learning (ML; Bishop, 2006) approaches, which are swiftly infiltrating many areas within healthcare and medicine for better informing and personalizing individual care (Deo, 2015; Johnson *et al.*, 2016a; Rajkomar *et al.*, 2019), have been introduced in the context of designing adaptive clinical trials. Given the sequential nature of such trials, a specific ML framework has emerged as a potential solution for solving the sequential, adaptive, learning problem. This is represented by the *reinforcement learning* (RL) framework (Sutton & Barto, 2018; Sugiyama, 2015).

Thus far, a plethora of methodological and theoretical studies have evaluated RL techniques in a variety of AIs, achieving performance exceeding that of alternative traditional techniques in many cases, and enhancing its use in real life.

Despite these propositions for the use of RL, and ML in general, to improve care delivery in medical research and practice, their practical use in (adaptive) clinical trials is still very limited. Nevertheless, the RL framework suggests to be particularly appropriate for clinical trials since the trade-off between clinical research and clinical practice can be seen as the well-known trade-off between exploration and exploitation. Under this pretext, development and evaluation of novel RL algorithms, specifically tailored to some of the above mentioned adaptive designs settings, is being increasingly interesting a broad literature. While, their introduction in the CTs arena can be traced back to the *multi-armed bandit* (MAB; Lattimore & Szepesvári, 2020) solution proposed by Thompson (1933), where with MAB we generally refer to a subclass of RL problems (as we will make it clear in Section 2), only recently, a sparkling interest in RL and MABs has emerged in relation to the CTs domain (Villar *et al.*, 2015a; Shen *et al.*, 2020). This trend is currently also being alimented by the increased attention of regulatory agencies for adaptive clinical trials (FDA, 2019; Pallmann *et al.*, 2018) and the unprecedented response in terms of clinical research activity of the COVID-19 (Zame *et al.*, 2020; Stallard *et al.*, 2020). Notably, the urgent need of sufficiently strong evidence required results to be obtained as rapidly

Table 3.3. Overview of common types of adaptive clinical trials designs and their characteristics

Adaptive Design	**Underlying Idea**	**Adaptive Element***
Dose-finding & seamless phase-I/II designs	Determine the most appropriate dose level in terms of either toxicity (*maximum tolerated dose*), or efficacy (*minimal effective dose*), or both (*seamless* phase-I/II trial designs), to be used in further phases	Treatment's dose
Sample size re-estimation (or N-adjustable) design	Adjusting or re-estimating the sample size to ensure the desired power	Total sample size
Adaptive randomization design	Skewing the randomization probabilities towards the most promising (e.g., with higher probability of success) or informative (e.g., that balance assignments within covariate profiles) treatment(s)	Randomization probability
Group sequential designs	Prematurely stopping the trial due to safety, efficacy or futility, according to a stopping criterion to either support or reject the sequential null hypotheses	Sample size (through the reduction in the number of groups due to a potential earlier stopping time)
Population-enrichment designs	Restricting future enrollment to subgroup(s) of patients more likely to benefit (most) from the treatment	Sample sizes of each group
Seamless phase-II/III design	Combining treatment selection (exploratory phase II: learning about the best treatment and dropping less efficacious or unsafe treatments) and efficacy confirmation (confirmatory phase III: testing hypothesis) into one trial	Compared treatments

*In each of the adaptive designs, in addition to the main adaptive element, other adaptations may be pre-specified, e.g., a seamless phase-II/III design may include an adaptive randomization probability as discussed in the illustrative example

as possible, making adaptive designs a particularly attractive option. However, existing heated debates around their reliable applicability in CTs argue for a deeper understanding of their mechanism and statistical properties.

In this work I conduct an extensive methodological review of current RL based methodologies proposed for designing adaptive CTs, with the aim of providing the research community with systematic understanding of theoretical foundations of this emerging paradigm, in conjunction with its applicability, potential benefits and

existing challenges within the CT domain.

Exclusively focusing on the types of adaptive CTs designs where the use of RL has been proposed, we show how these can be naturally formalized through the RL framework, and discuss their key differences under this common framework. With this unified understanding we hope to offer a foundation to more easily conduct research in both theoretical and applied sciences. Subsequently, we separately tackle the above mentioned adaptive designs (more specifically, adaptive-dose finding, group sequential testing and adaptive randomization), reviewing the proposed RL technique in each domain and discussing the related benefits and open problems.

Please note that this review (unlike the tentative approach taken for the adaptive intervention setting) is not meant as a comprehensive a fully-inclusive work of all the existing settings and methods. Just to mention, I identified an additional area in the design of adaptive clinical trials which used some form of RL for adaptively improving the design of the study. It is the case of *group sequential tests*, which deal with *repeated significance testing* applied at predefined time points, or after having sampled a group of observations, to cumulating data, with critical boundaries adjusted for multiple testing. We refer to Jennison & Turnbull (2000) and Proschan *et al.* (2006) for an overview and general concepts of the statistical methodology in group sequential designs; and point to Jennison & Turnbull (2013) for a RL-technique used in this setting (the only existing one, to our knowledge).

The focus of the next two sections will be on the the design of adaptive-dose finding trials and response-adaptive randomization, and existing RL-based methodologies proposed in these areas. Connecting the RL, MAB, and clinical trials design literature, we report in Table 3.4 the correspondence between these literatures' terminologies.

Table 3.4. Reference terminology in reinforcement learning (RL), multi-armed bandits (MABs) and Dose-finding and Response-adaptive randomization (RAR) designs

Notation	**Terminology**			
	RL	MABs	Dose-finding	RAR
i	Trajectory	Trajectory	Trial*	Trial
t	Time	Round	t-patient/t-cohort*	t-patient/t-cohort
X	State	Context	Covariates	Covariates
A	Action	Arm	Treatment dose	Treatment
Y	Reward	Reward	Toxicity (+Efficacy)	Outcome/Response
$\mathbf{H}$	History	Filtration	Past observations	Observations sequence
$\boldsymbol{\pi}/\mathbf{d}$	Policy	Policy	Dosage allocation	Allocation sequence

*In (adaptive) dose-finding designs, sometimes i is also used to indicate a single patient or cohort, which is followed for t times.

3.2.1 Adaptive Dose-Finding Designs

An adaptive dose-finding design is often used in early-phases (I or I-II) clinical trials development to determine the most appropriate dose level that should be used in further phases of the trial (Yuan *et al.*, 2017). Conventional phase I designs focus on determining the highest dose with acceptable toxicity, called the *Maximum Tolerated Dose* (MTD). More specifically, given K different dose levels that have been chosen

by physicians based on preliminary experiments (K is usually a number between 3 and 10), and denoting by p_k the (unknown) toxicity probability of dose k, the MTD is defined as the dose with a toxicity probability closest to a target:

$$\text{MTD} = k^* \in \underset{k \in \{1,\dots,K\}}{\arg\min} |p_k - \theta|, \tag{3.24}$$

where θ is the pre-specified targeted toxicity level (TTL), which is determined by clinical expertise, evidence from previous studies, and guidance from the trial statistician, and is typically set between 0.2 and 0.35 (Wheeler *et al.*, 2019).

Once an MTD has been determined, and safety is no longer a major concern, early efficacy is evaluated in phase II trials on larger groups to evaluate whether an experimental treatment is promising (Yuan *et al.*, 2017; O'Quigley *et al.*, 1990). For clinical trials in life-threatening diseases, efficacy is often assumed to be increasing with toxicity and dose, hence the MTD is the most appropriate dose to further investigate in the rest of the trial. However, the assumption that both toxicity and efficacy of the treatment are monotonically increasing with the dose (Chevret, 2006), has been shown to not hold in general (Riviere *et al.*, 2018). Thus, recently, more complex designs evaluating simultaneously both toxicity and efficacy have been considered in dose-finding methods to accelerate the development process of new treatments and to reduce costs (Thall & Cook, 2004; Zhang *et al.*, 2006; Zang *et al.*, 2014; Riviere *et al.*, 2018). These "multi objective" dose-finding designs, known as "seamless" phase I-II trial designs (Mahajan & Gupta, 2010), aim to estimate the smallest dose to obtain a desired efficacy, called the *Minimal Effective Dose* (MED), while satisfying specific MTD safety requirements. Because little is known about the new drug in the early phase investigation, these studies are naturally conducted in an adaptive and small-group-sequential manner, characterized by an iterative process.

A general overview of the dose-finding problem and related adaptive designs can be found in Chevret (2006); Yin (2012), while a more up-to-date review of modern methods and theory in Cheung (2015). They discuss the widely used rule-based approaches (O'Quigley & Zohar, 2006), model-based designs such as the Continual Reassessment Method (CRM; O'Quigley *et al.*, 1990; Wheeler *et al.*, 2019) and other Bayesian strategies (Yin *et al.*, 2006; Yin, 2012), and solve the adaptive problem in a traditional statistical way, as opposed to RL-based approaches.

Here, given the general aim of this review work, we focus our attention on the existing methods in dose-finding trials designs which used novel ideas based on RL, which include both MAB-based strategies such as Thompson Sampling and Upper Confidence Bound (see Algorithms 4 and 3) and Q-learning (see Algorithm 1).

Thompson Sampling for MTD Identification Interestingly, in the growing literature on Bayesian adaptive designs (Berry *et al.*, 2010), several designs that may be viewed as variants of Thompson Sampling have been proposed (Thall & Wathen, 2007; Satlin *et al.*, 2016). However, to the best of our knowledge, only very recently, the use of MAB algorithms, more precisely Thompson Sampling, has been investigated for dose-finding trials (Aziz *et al.*, 2019). The proposed MAB approach by Aziz *et al.* (2019) is based on the well-known exploration-exploitation problem: finding the MTD (which is crucial for the next stages of the trial) and treating as many trial

participants as possible with this MTD, which is a common issue in clinical trials. By viewing optimal dose identification as a particular multi-armed bandit problem, authors rephrase the exploration (finding the best treatment dose, i.e., MTD) and exploitation (treating as many trial participants as possible with this MTD) trade-off as a trade-off between error probability and rewards, two performance measures that are well-studied in the bandit literature and that are known to be somewhat antagonistic (Bubeck *et al.*, 2011; Kaufmann & Garivier, 2017).

They propose a simple statistical model for the MTD identification problem in phase I clinical trials, and show that it can be viewed as a particular multi-armed bandit problem, with the notion of optimal arm naturally defined as the arm with mean closest to the TTL threshold θ in the MTD identification problem, as in (3.24). We remind that a MAB model refers to a situation in which an agent sequentially chooses arms (here doses) and gets to observe a reward (here a realization of an underlying probability distribution which characterises the probability that the chosen dose is toxic).

Specifically, denoted with A_t the dose at round t, for each t, the MAB strategy selects and administers a dose $A_t \in \{1, \dots, K\}$ to a patient for whom a toxicity response $Y_{t+1}^{\text{tox}}(A_t)$ is then observed. Assuming a binary reward variable $Y_{t+1}^{\text{tox}}(A_t)$, with $Y_{t+1}^{\text{tox}}(A_t) = 1$ indicating that a harmful side-effect occurred and $Y_{t+1}^{\text{tox}}(A_t) = 0$ than no harmful side-effect is present, $Y_{t+1}^{\text{tox}}(A_t)$ is modelled by a Bernoulli distribution with mean p_{A_t}. Assuming, now, a prior distribution over the vector of toxicity probabilities $\boldsymbol{p} = (p_1, \dots, p_K) \in [0,1]^K$, e.g., $\Pi_0 = \prod_{k=1}^K \pi_{k,0}$ where each $\pi_{k,0} = \mathcal{U}[0,1]$ is an independent uniform distribution, authors advocate the use of the *Independent TS* for MTD identification as follows. For each time t, a sample $\tilde{\boldsymbol{\theta}}_t = (\tilde{\theta}_{1,t}, \dots, \tilde{\theta}_{K,t})$ from the posterior distribution $\Pi_t = \prod_{k=1}^K \pi_{k,t}$ of the toxicity probability vector $\boldsymbol{p}$ is generated for each dose k, and the dose for which the sample is closest to the TTL θ is selected.

More formally, for the given prior and likelihood, at each time t and $\forall k \in \{1, \dots, K\}$,

$$\begin{aligned} \tilde{\theta}_{k,t} &\sim \pi_{k,t} = \text{Beta}(S_{k,t} + 1, N_{k,t} - S_{k,t} + 1), \\ A_{t+1} &\in \underset{k \in \{1,\dots,K\}}{\arg\min} |\tilde{\theta}_{k,t} - \theta|, \end{aligned} \tag{3.25}$$

where $\text{Beta}(S_{k,t} + 1, N_{k,t} - S_{k,t} + 1)$ represents the conjugate posterior distribution (note that the Uniform prior is equivalent to a $\text{Beta}(1,1)$), with $S_{k,t} = \sum_{s=1}^t Y_s^{\text{tox}} \mathbb{I}_{\{A_s = k\}}$ the number of times a harmful toxicity from dose k has been registered, $N_{k,t} = \sum_{s=1}^t \mathbb{I}_{\{A_s = k\}}$ the number of times dose k has been given, and A_s is the dose allocated at time s.

As the TS randomization induces some exploration, and recommending $\hat{k}_t = A_{t+1}$ might not be the best idea, authors propose the idea of recommending $\hat{k}_t = \arg\min_k |\hat{\mu}_{k,t} - \theta|$, where $\hat{\mu}_{k,t}$ is the empirical mean of dose k after the t-th patient of the study as in Bubeck *et al.* (2011) or $\hat{k}_t = \arg\max_k N_{k,t}$.

Please note that this is a simplified version, in the sense that it is a context-free strategy, of the TS with linear reward algorithm introduced in Section 3.1.2. In addition, by using a Beta-Bernoulli model, it has a directly interpretable objective function in terms of mean toxicity probabilities $p_{k,t}$'s, or $\tilde{\theta}_{k,t}$'s.

As a further development, authors also show that using more sophisticated prior distributions allows the algorithm to leverage some particular constraints of the dose-finding problem, like increasing toxicities or a plateau of efficacy, considering a multi-outcome optimization problem.

For example, by assuming a two-parameter Bayesian logistic model (that is among the most popular - also used in the CRM) for increasing toxicity probabilities, i.e.,

$$p_k(\beta_0, \beta_1) = \frac{1}{1 + e^{-\beta_0 - \beta_1 u_k}}, \tag{3.26}$$

with u_k the *effective dose*, and β_0 and β_1 the model's parameters for which $\mathcal{N}(0, 100)$ and $\mathcal{E}(1)$ priors are considered, respectively, the following TS strategy can be adopted:

$$\begin{aligned} \left(\tilde{\beta}_0, \tilde{\beta}_1\right)_t &\sim \pi_t, \\ A_{t+1} &\in \underset{k \in \{1,\dots,K\}}{\arg\min} \left| p_k(\tilde{\beta}_{0k,t}, \tilde{\beta}_{1k,t}) - \theta \right|. \end{aligned}$$

Posterior distribution π_t over the parameters (β_0, β_1) at time t is not available in closed form, and it can be approximated by using Hamiltonian Monte Carlo (HMC) Markov Chain algorithms such as in Aziz *et al.* (2019).

As for the *Independent TS* in (3.25), authors do not recommend the use of $\hat{k}_t = A_{t+1}$ as MTD, but instead $\hat{k}_t \in \arg\min_{k \in \{1,\dots,K\}} \left| p_k(\hat{\beta}_{0k,t}, \hat{\beta}_{1k,t}) - \theta \right|$, with $\hat{\beta}_{0k,t}$ and $\hat{\beta}_{1k,t}$ the posterior means, typically considered in the CRM literature.

Finally, in order to take into account both toxicity constraints and dose effectiveness, the notion of *admissible set of doses* $\mathcal{A}_t$, inspired by Riviere *et al.* (2018) and described in detail in Aziz *et al.* (2019), is introduced, and the optimization problem focuses on finding the MED. Practically, each time a not admissible candidate dose $A_t \notin \mathcal{A}_t$ is given by sampling from the posterior, the sampling process has to be repeated again, limiting the exploration of the TS; while in the second case, the optimization problem for the optimal dose considers a two dimensional reward vector $(Y^{\text{tox}}_{k,t}, Y^{\text{eff}}_{k,t})$, with $Y^{\text{eff}}_{k,t}$ an efficacy outcome, and an additional parameter q_k denoting the efficacy probability of dose k. Thus, if we now consider as admissible set $\mathcal{A}_t = \{k : p_k \leq \theta\}$, the optimal dose finding problem becomes

$$k^* = \min \left\{ k : q_k = \max_{l \in \mathcal{A}_l} q_l \right\}. \tag{3.27}$$

Following Riviere *et al.* (2018), again a model based approach is considered: toxicity follows the two-dimensional Bayesian logistic model presented in (3.26) and efficacy also follows a logistic model, with an additional parameter τ that indicates the beginning of the plateau of efficacy. Efficacy and toxicity are assumed to be independent. Based, on this model, TS is carried out in the general way, with samples this time from the multivariate posterior distribution, obtained with the HMC technique for the continuous parameters of the logistic functions and the random sampling from the conditional distribution of the discrete parameter τ.

Contextual Constrained Learning: a UCB-based Approach Following the multi-criteria dose-finding strategy of Aziz *et al.* (2019) presented in (3.27), and the multi-dimensional toxicity and efficacy outcomes vector, an alternative bandit

solution, based on the UCB principle, is developed in Lee *et al.* (2020). They propose the *C3T-Budget*, a solving strategy for what they call *contextual constrained clinical trial* (C3T) problem for dose-finding. In the C3T problem, patients arrive sequentially and the agent has to determine which patients to treat and the dose to be allocated to the patient, based on both budget (maximum number of patients that can be admitted into the trial) and safety constraints, also considering heterogeneous groups of patients, that makes the dose-finding problem more complex (Wages *et al.*, 2015).

The proposed C3T-Budget algorithm selects at each time t the dose $A_{s,t} \in \mathcal{A} = 1, \dots, K$ for each group s so as to maximize the expected efficacy (*clinical practice*), while satisfying the safety constraint based on information learned from previously treated patients. Then, given the chosen dose, the algorithm determines whether the patient is treated or not (we denote with $k = 0$ the absence of treatment) with the aim of maximizing the information from *clinical research*, balancing thus the trade-offs between information gathering and treatment effectiveness.

Formally, denoted with $s \in \mathcal{S} = \{1, \dots, S\}$, the observed subgroup over a time-horizon T, with B the limited budget, and with $Y_{s,k}^{\text{tox}}$ and $Y_{s,k}^{\text{eff}}$ the toxicity and the efficacy outcomes of dose k for subgroup s, authors model the outcomes as Bernoulli random variables with unknown parameters $p_{s,k}$ and $q_{s,k}$, respectively. $Y_{s,k}^{\text{tox}} = 1$ indicates that dose k is unsafe for subgroup s, and $Y_{s,k}^{\text{eff}} = 1$ that dose k is effective for subgroup s. We remind that a dose k is considered unsafe if the expected toxicity $p_{s,k}$ exceeds the MTD threshold (or TTL) θ; it is considered ineffective if a minimum efficacy threshold ψ is not reached. For $p_{s,k}$, is considered again the logistic dose-toxicity model $p_{s,k}(a) = (\frac{\tanh u_k + 1}{2})^a$, as in O'Quigley *et al.* (1990), where u_k is the effective dose level of dose k and a a common dose parameter.

The cost/budget variable at time t is denoted by Y_t^{cost}, and it adds one unit each time a patient is administered a dose (i.e., $A_t = k \neq 0$); in this case we have $Y_t^{\text{tox}} = Y_t^{\text{tox}}(A_t, X_t)$, $Y_t^{\text{eff}} = Y_t^{\text{eff}}(A_t, X_t)$ and $Y_t^{\text{cost}} = 1$, with $X_t \in \mathcal{S}$ representing the contextual subgroup variable. When $A_t = 0$, we have no efficacy or toxicity and the cost $Y_t^{\text{cost}} = 0$. The clinical trial ends when the budget is exhausted or at the end of time-horizon T.

To make a recommendation, the MTD threshold and minimum efficacy threshold are considered: by defining the set of *candidate doses* (analogously to admissible set) for subgroup s by $\mathcal{K}_s = \{k : q_{s,k} \geq \psi, p_{s,k} \leq \theta\}$, the optimal dose-to-recommend for each subgroup s is given, similarly to (3.27), by:

$$k_s^* = \begin{cases} \arg\max_{k \in \mathcal{K}_s} q_{s,k}, & \text{if } \mathcal{K}_s \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Denoting now with $\hat{k}_s^*(T, B)$ the estimated dose to recommend to subgroup s at the end of the clinical trial, and incorporating the budget constraint, the C3T problem becomes a problem of minimizing the dose recommendation error while satisfying

the budget and safety constraints, and it is formally presented as

$$\begin{aligned} \text{minimize} \quad & \sum_{s \in S} \mathbb{E}\Big[\mathbb{I}[\hat{k}_s^*(T,B) \neq k^*]\Big] \\ \text{subject to} \quad & \mathbb{P}\Big[\overline{p}_s(T,B) \leq \theta\Big] \geq 1-\delta_s, \quad \forall s \in \mathcal{S} \\ & \sum_{t=1}^{T} Y_t^{\text{cost}} \leq B, \end{aligned}$$

where $\overline{p}_s(T,B) = \frac{\sum_{t=1}^{T} \mathbb{I}[X_t=s] Y_t^{\text{tox}}}{\sum_{t=1}^{T} \mathbb{I}[X_t=s]\mathbb{I}[A_t \neq 0]}$ is the expected toxicity of subgroup s, and δ_s the maximum probability with which the toxicity for subgroup s can exceed the TTL threshold θ.

For solving this minimization problem, at each round t, the *C3T-Budget* first, constructs the sets of candidate doses for s, $\mathcal{K}_s$, by considering the estimated expected efficacy and toxicity of each dose for the subgroup. Then, among the candidate doses, the proposed algorithm selects the estimated optimal dose at round t, $k_{s,t}^*$, that has the largest UCB of the expected efficacy for subgroup s, i.e., $k_{s,t}^* = \arg\max_{k \in \mathcal{K}_s} \hat{q}_{s,k,t}$. Here, $\hat{q}_{s,k,t}$ denotes the UCB of $q_{s,k}$ at time t, and is given by $\hat{q}_{s,k,t} = \overline{q}_{s,k} + \sqrt{\frac{c \log N_{s,t}}{N_{s,k,t-1}}}$, with $N_{s,t} = \sum_{\tau=1}^{t} \mathbb{I}[X_\tau = s]$ the number of times subgroup s has arrived up to time t, $N_{s,k,t-1} = \sum_{\tau=1}^{t-1} \mathbb{I}[X_\tau = s, A_\tau = k]$ the number of times that dose k has been allocated to subgroup s up to round $t-1$, and $\overline{q}_{s,k} = \frac{\sum_{\tau=1}^{t} \mathbb{I}[X_\tau=s, A_\tau=k] Y_t^{\text{eff}}}{N_{s,k,t}}$ the empirical efficacy estimation of dose k. Finally, the agent determines whether the patient in round t is to be skipped or not by considering how convincing the estimation of the efficacy of k_s^* is. To do this, the Bayesian credible interval of the estimation of $\overline{q}_{s,k}$ is adopted.

In a simulation study, authors compare their proposed *C3T-Budget* method to the Contextual *Independent TS* of Aziz *et al.* (2019) introduced in (3.2.1), as well as the Contextual UCB introduced in Section 3.1.2 and some rule-based standard techniques, showing its out-performance in terms of both total error rate and efficacy per patient.

Based on the same UCB principle, a more recent strategy is illustrated in Shen *et al.* (2020), who proposed the *Safe Efficacy Exploration Dose Allocation* (SEEDA) algorithm. This novel adaptive clinical trial methodology, explicitly imposes safety constraints to the allocation and recommendation of dose levels, while maximizing the cumulative efficacies, by adaptively updating the admissible set of dose levels with UCB. Experiments on simulated datasets, as well as clinical trials built from real-world datasets, show that the proposed method is capable of finding the optimal dose with higher success rate and fewer patients, compared to other state-of-the-art designs (Shen *et al.*, 2020).

Approximate Dynamic Programming for Hybrid Designs Optimization
A different strategy is considered in Bartroff & Lai (2010), where, again the optimal phase I design is formulated as in Aziz *et al.* (2019) and Lee *et al.* (2020) so as to incorporate the *treatment versus experimentation dilemma*, addressing the ethical

issue of treating patients at dose levels below the unknown MTD for safety, and hopefully close to the MTD for efficacy. Here, authors formulate the problem as a stochastic optimization problem, and, by making use of recent advances in *approximate dynamic programming* (ADP), develop a new tool for tackling the optimization problem and derive an approximated optimal design in a Bayesian fashion.

The proposed design, is a presented as a convex combination, therefore *hybrid design*, of a "treatment" design that is targeted toward treating the current patient at the best guess of the MTD, and a "learning design", aiming to efficiently experiment and gather information for future patients. Following the hybrid idea, and denoting by a_i, $i = 1, \ldots, N$, the dose administered to patient i, a representation of the optimal dose sequence is given by

$$a_i^* = (1 - \epsilon_i)m_i + \epsilon_i l_i, \tag{3.28}$$

where l_i is the chosen "learning design" and m_i the myopic dose that minimizes a risk (toxicity) function. The weights ϵ_i's in these convex combinations are determined by ADP, more specifically by using *rollouts*, and can be conveniently stored to provide simple table look-up schemes for the clinical user.

We remind that *dynamic programming* (DP) is a standard optimization approach to a general stochastic optimization problem of the form

$$\mathbb{E}\Big[\sum_{i=n}^{N} h(a_i, \eta) + g(\hat{\eta}, \eta)|\mathbf{h}_N\Big] = \mathbb{E}\Big[h(a_n, \eta) + \sum_{i=n+1}^{N} h(a_i, \eta) + g(\hat{\eta}, \eta)|\mathbf{h}_N\Big], \tag{3.29}$$

based on which, a simplification acts upon the complicated decision by breaking it down into a sequence of simpler sub-decision in a recursive manner, and a solution is given by using for instance the *Bellman equation* (see Equation (2.10) in Section 2). In (3.29) the expectation is taken over the joint distribution of $(\eta; A_1, Y_1, \ldots, A_N, Y_N)$, with Y_i the outcome for patient i, $\mathbf{h_i} = (a_1, y_1, \ldots, a_i, y_i)$ the "history", or the information set generated by the first i doses and responses, and η the MTD. h and g are two functions quantifying the risk of toxicity in relation to dose a_i, and more generally they denote an outcome/reward function. The *state-value function* in (2.5) and *action-value function* in (2.6) are typical examples of (3.29).

To minimize the *global risk* in (3.29), dynamic programming solves for the optimal design $a_1^*, \ldots, a_N^*$ by backward induction that determines a_i^* after determining the future dose levels $a_{i+1}^*, \ldots, a_N^*$. It involves computing and minimizing the conditional expectations over all a, which makes it a formidable hard task.

To overcome this difficulty, approximated DP solutions which go under the rubric of ADP, and combine least squares regression with Monte Carlo simulations, have been employed. However, rather than having an *approximation in the value space*, as seen in Section 3.1.1 for Q-learning, authors propose *approximating the policy space*, which uses iterated *rollouts* to optimize the parameters in a suitably chosen parametric family of (dosing) policies. The choice of the family of policies should involve domain knowledge and reflect the kind of policies that one would like to use for the actual application. The main idea behind the rollout is iterative policy

improvement: starting with a base policy $\hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_N)$, the rollout of $\hat{\mathbf{a}}$, i.e., $\mathbf{a}^{(1)} = (\hat{a}_1^{(1)}, \ldots, \hat{a}_N^{(1)})$ is given by the dose that minimizes a function similar to the one in (3.29); then $\mathbf{a}^{(1)}$ can be used as base policy for further rollouts yelding the optimal design after n number of times. Further details of this approach can be found in Bartroff & Lai (2010), Bartroff *et al.* (2013) and Chapter 6 of Bertsekas (2011).

Note that (3.29) measures the effect of the dose a_n on the n-th patient through $h(a_n, \eta)$, its effect on future patients in the trial through $\sum_{i=n+1}^{N} h(a_i, \eta)$, and its effect on the post-trial estimate $\hat{\eta}$ through $g(\hat{\eta}, \eta)$. It can therefore be used to address the dilemma between safe treatment and experimentation and can be viewed as the hybrid design in (3.28) by relating the myopic dose m_n to the minimum of the proximal risk function $h(a_n, \eta)$ - if the n-th patient were the last patient to be treated in the trial ($n = N$), and the l_n design to the perturbation one expects to have from m_n in the direction of a dose that provides more information about the dose-response model, for the relatively large number of doses that will have to be set for the future patients, i.e, $\sum_{i=n+1}^{N} h(a_i, \eta)$. Since the trade-off quantified by the weights in (3.28) stems from the uncertainty in the current estimate of the MTD η (see Bartroff & Lai, 2010; Bartroff *et al.*, 2013, for more details), it is natural to expect that the amount of perturbation from the myopic dose m_n depends on the degree of such uncertainty, using little perturbation when the posterior distribution of η (here again a two-logistic model as in the CRM is assumed) is peaked, and much more perturbation when it is spread out. This suggests choosing ϵ_i's as functions of the posterior variance, such as $\epsilon_i = \epsilon_i(s_i)$, with $s_i = v_{i-1}/v_0$ (basic features of the posterior distribution of η used to approximate the ϵ_i) and $v_{i-1}^2 = Var(\eta|\mathbf{h}_{i-1})$. The ADP rollout algorithm is used to determine the functions $\epsilon_i = \epsilon_i(s_i)$.

Q-learning: dose-finding as a DTR identification problem We finally want to point out that an adaptive dose-finding problem can be actually regarded as a sequential decision-making problem in which different treatments consist in different doses. Based on this relationship, in such phase I/I-II trials, the attending physician may actually use a DTR for making personalized multi-cycle decisions for each patient. Depending on the patient's history of doses and outcomes, the dose given in each cycle may be above, below, or the same as the dose given previously, or therapy may be terminated due to excessive toxicity or poor efficacy. Since typical early-phase trial designs ignore such within-patient multi-cycle decision making, the "optimal" dose chosen by such a design actually pertains only to the first cycle of therapy.

The problem of adaptively optimizing each patient's doses (given in multiple cycles) based on binary efficacy and toxicity has been firstly addressed in Lee *et al.* (2015a). Authors employ a model-based Bayesian objective function, defined in terms of efficacy and toxicity utilities, which structurally resembles Q-learning functions (Watkins, 1989). The goal of the proposed method is to choose a dose in each cycle so as to maximize the posterior expected mean of the objective function, applying a modified recursive *Bellman equation* (see Equation (2.10)). More specifically, denoting the set of actions or doses with $\mathcal{A} = \{0, 1, \ldots, K\}$, where $A = 0$ is the decision to not give the patient any treatment, with $Y_{i,t}^{\text{tox}} \in \{0, 1\}$ and $Y_{i,t}^{\text{eff}} \in \{0, 1\}$

the outcome indicators for toxicity and efficacy of patient i at cycle t, the mean utility, or mean reward function, of action a_2 in cycle $t = 2$ is defined as

$$\begin{aligned} Q_2(a_2, a_1, Y_1^{\text{tox}}, Y_1^{\text{eff}}, \theta) &= \mathbb{E}[U(Y_2^{\text{tox}}, Y_2^{\text{eff}})|a_2, a_1, Y_1^{\text{tox}}, Y_1^{\text{eff}}, \theta] \\ &= \sum_{y_2^{\text{tox}}=0}^{1} \sum_{y_2^{\text{eff}}=0}^{1} U(y_2^{\text{tox}}, y_2^{\text{eff}}) p(y_2^{\text{tox}}, y_2^{\text{eff}}|a_2, a_1, Y_1^{\text{tox}}, Y_1^{\text{eff}}, \theta), \end{aligned}$$

with U reflecting patients' utility, e.g., $U(0, 1) = 100$ and $U(1, 0) = 0$, and p the joint likelihood for the observables of a patient.

The objective function at cycle 2 is the posterior expected utility of giving dose a_2 in cycle 2 defined as

$$q_2(a_2, a_1, Y_1^{\text{tox}}, Y_1^{\text{eff}}, \mathcal{D}_n)) = \mathbb{E}[Q_2(a_2, a_1, Y_1^{\text{tox}}, Y_1^{\text{eff}}, \theta|a_2, a_1, Y_1^{\text{tox}}, Y_1^{\text{eff}}, \mathcal{D}_n)],$$

where $\mathcal{D}_n)$ indicates current data, and an optimal dose is given by the one which maximises the given objective. The dose-finding process proceeds backward for the optimal stage 1 dose as illustrated in Algorithm (1).

Since maximizing a posterior utility-based objective function, *per se*, ignores the undesirable but important possibility that considered options are too toxic, *dose acceptability* constraints criteria are included in the algorithm. More specifically, denoted by a_1^M and a_2^M the highest doses among those that have been tried in cycle 1 and either cycle respectively, the search for optimal actions is constrained so that $1 \leq a_1 \leq \min(a_1^M + 1, K)$ and $1 \leq a_2 \leq \min(a_2^M + 1, K)$: the first constraint is that an untried dose level may not be skipped when escalating; the second constraint does not allow escalating a patient's dose in cycle 2 if toxicity was observed in cycle 1, $Y_1^{\text{tox}} = 1$; the third criterion, defined in terms of expected utility, is to avoid giving undesirable dose pairs.

The overall method provides an optimal two-stage regime consisting of an optimal cycle 1 dose, and an optimal function of the patient's cycle 1 dose.

A similar sequentially outcome-adaptive Bayesian design, using utilities of a bivariate (toxicity and efficacy), but ordinal, outcome is proposed also in Thall & Nguyen (2012). However, rather than choosing the dose that maximizes the posterior mean utility, they deal with the well-known "exploration versus exploitation" dilemma (Sutton & Barto, 2018) by adaptively randomizing patients among a set of *acceptable doses*, $\mathcal{A}_n$. In this case, a dose is considered acceptable if it (i) has acceptable toxicity, (ii) has posterior mean utility that is close to the maximum, and (iii) is not unlikely to have the highest posterior utility. In addition, in order to define the adaptive randomization (AR) probabilities, the acceptable dose set $\mathcal{A}_n$ is also refined in relation to a set of "good" outcomes, defined as $G = \{(y^{\text{tox}}, y^{\text{eff}}) : U(y^{\text{tox}}, y^{\text{eff}}) \geq \underline{U}\}$, where the lower limit $\underline{U}$ is elicited from the physicians who provided the utilities. Now, given G, and denoting the posterior mean $\mu_G(a, \mathcal{D}_n) = \mathbb{E}[\mathbb{P}((Y^{\text{tox}}, Y^{\text{eff}}) \in G|a, \mathcal{D}_n)]$, where $\mathbb{P}((Y^{\text{tox}}, Y^{\text{eff}}) \in G|a, \mathcal{D}_n)$ is the *probability of a good outcome for a patient treated with dose* a, the idea is to randomize a patient to dose $a \in \mathcal{A}$ with probability

$$\pi(a, \mathcal{D}_n) = \frac{\mu_G(a, \mathcal{D}_n)}{\sum_{z \in \mathcal{A}_n} \mu_G(z, \mathcal{D}_n)}.$$

Adaptively randomizing patients in clinical trials is of increasing use in clinical practice, and the problem is increasingly studied in clinical literature. We specifically address this problem, reviewing existing RL and MAB-based techniques in Section 3.2.2; however, as in Thall & Nguyen (2012) AR is used as part of the primary objective of dose-finding, we decided to mention this work in the current section.

3.2.2 Response-Adaptive Randomization

RCTs have traditionally followed a static design in which patient allocation to treatments is fixed throughout the trial, typically based on a *uniform randomization* or a *blocked randomization* (Lachin *et al.*, 1988), to prevent an imbalance between the groups. The primary goal of this static design is to learn about the efficacy of treatments and allow comparison between the treatment(s) and the control. Adaptive randomization (AR) designs, where assignment to treatments evolves as patient outcomes are observed, are gaining in popularity due to potential for improvements in cost and efficiency over traditional designs. They use data of previous cohorts to adapt allocation of patients in succeeding cohorts: if a particular treatment showed more promising or informative results in prior patients, the probability of being assigned to that treatment is increased (Hu & Rosenberger, 2006; Berry, 2006). Commonly used AR procedures include restricted or treatment-adaptive randomization (Efron, 1971b; Wei, 1978), covariate-adaptive randomization (Zelen, 1974; Taves, 1974), and response-adaptive randomization (RAR; Rosenberger & Lachin, 1993; Hu & Rosenberger, 2006). An extensive overview of these AR designs can be found in Rosenberger & Lachin (2015) and Antognini & Giovagnoli (2015). Here, we focus on RAR, where the idea is to skew the sequential allocation procedure in favour of the treatments associated with the best response (including also historical allocation or covariate knowledge, if any). The goal is to improve the overall benefit to patients. However, at the same time, we also require some allocation of the worse treatments which will enable us to make meaningful inferences about treatment differences or other parametric functions of interest. Thus, such adaptive strategies should lead to allocating a larger number of patients to the eventual better treatment, without significantly weakening the strength of the comparison between treatments.

RAR methods for CTs have been studied by many authors and have a long history within CTs literature. Traditional statistical contributions include non-parametric *procedures based on urn models*, such as the pioneering concept of *Play-the-Winner* (PW; Zelen, 1969; Robbins, 1952; Hoel & Sobel, 1971; Wei & Durham, 1978) rule, *generalized urn designs* (Athreya & Karlin, 1968; Durham & Yu, 1990; Durham *et al.*, 1998), *birth and death urn designs* (Ivanova *et al.*, 2000) or *Drop-the-Loser* (DL; Ivanova & Durham, 2000; Ivanova, 2003; Zhang *et al.*, 2007) rules, and parametric *procedures based on sequential estimation*, such as the *doubly-adaptive biased coin designs* (Eisele, 1994; Eisele & Woodroofe, 1995; Hu *et al.*, 2004). As noticed by Hu & Rosenberger (2006), these procedures are *myopic* or *one-step-look-ahead*, in that they incorporate current data on treatment assignments and responses into decisions about treatment assignments for the next subject only. Thus, there is no guarantee that such procedures are globally optimal.

A different line of research, which aims to find an allocation sequence (i.e., policy) with the objective of maximizing the cumulative outcomes (patient benefits)

over the total sample or horizon, is based on bandit problems (Berry & Fristedt, 1985a; Gittins *et al.*, 2011; Villar *et al.*, 2015a). This line has an old and long history in statistics, starting with the Bayesian proposal of Thompson (1933, 1935). However, after his pioneering work, the bandit problem received little attention until the works of Robbins (1952); Bellman (1956) and, later, the celebrated results of Gittins (1974) for Bayesian bandits. Indeed, much of the bandit literature in ADs takes the Bayesian approach: as mentioned in Berry & Fristedt (1985a), *"It is not that researchers in bandit problems tend to be 'Bayesians'; rather, Bayes's theorem provides a convenient mathematical formalism that allows for adaptive learning, and so is an ideal tool in sequential decision problems"*. With a Bayesian strategy, a bandit is a typical DP problem; when the horizon is finite, backward induction can be used to determine an optimal allocation. Berry & Fristedt (1985a) also provides a comprehensive overview and survey of all the early bandits-based methods for adaptive randomization (see the *"Annotated bibliography"* pp.207-261). Here, we aim to focus on the more recent literature, reviewing also some of the early methods when necessary. The essential criterion for inclusion is that they resemble the RL or MAB problem, i.e., directly or implicitly, maximizing a sum (here we only deal with the discrete case), perhaps with discounting.

We now introduce some basic notation. In RAR designs, we wish to allocate a total of N patients to $K+1$ available treatments (say $A_0, \dots, A_K$, with A_0 denoting the control treatment), with patients recruited in a total of T time steps. Patients can arise sequentially in cohorts, in which case $T < N$, or one after the other at subsequent steps. In the latter, each time step t will be uniquely identified by the n patient entering the trial, thus $N = T$. We focus for now on this case. Each patient t of our sample will be assigned one of the available treatments and a response Y_t will be observed before making the treatment decision for patient $t+1$. The objective is to find an allocation rule so as to maximize the expected discounted return given in (2.3). Formally, denoted with γ the discount rate, and $\mathbf{a_t} \doteq (a_{0,t}, \dots, a_{K,t})$ the allocation vector for patient t, where $a_{k,t} = 1$ if unit t is allocated to treatment k and $a_{k,t} = 0$ otherwise, we want to choose the optimal allocation strategy $\boldsymbol{\pi}^* \doteq \{\pi_t^*\}_{t=1}^T$ so that

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=1}^{T} \gamma^t \sum_{k=0}^{K} a_{k,t} Y_t \right], \tag{3.30}$$

where the expectation $\mathbb{E}_{\boldsymbol{\pi}} = \mathbb{E}_{P_{\boldsymbol{\pi}}}$ is based on the entire trajectory distribution as in (2.2), with policy $\boldsymbol{\pi}$, and $\boldsymbol{\Pi} = \prod_{t=1}^{T} \Pi_t$, with $\Pi_t \subset \{0,1\}^{K+1}$ is the family of admissible policies, that is, all the feasible sequences of treatment allocations $a_{k,t}$ for all k and t. Clearly, given that only one treatment can be allocated per patient we have that $\sum_{k=0}^{K} a_{k,t} = 1$ for each t. The reward variable Y_t at time t is a function of both the treatment A_t and the state X_t, which can incorporate covariate information on that patient, but also previous responses information; we alternatively denote it with $Y_{k,t}$, indicating the response to treatment k for patient t.

If we consider the example of random and binary responses, that is, reward is either a success ($Y_{k,t} = 1$) or a failure ($Y_{k,t} = 0$), and model the sequence of successes and failures as a Bernoulli process, with $p_k = \mathbb{P}(Y_{k,t} = 1)$ the true unknown success probability of arm k, then, each treatment may be modelled as a *Markovian bandit*

process with state $X_{k,t}$ the random vector of successes $S_{k,t}$ and failures $F_{k,t}$ up to time t, i.e., $X_{k,t} \doteq (S_{k,t}, F_{k,t})$; it can be interpreted as a *sufficient statistic* for arm k. States update are based on the Bayes rule (with a Beta prior, with positive constant hyperparameters $s_{k,0}, f_{k,0}$), and the reward function is the average reward given by the mean of the Beta function, i.e., $Y_{k,t} = \mathbb{E}[p_k|X_{k,t}] = S_{k,t}/\left(S_{k,t} + F_{k,t}\right)$. This setting is also known as the *Bayesian Bernoulli MAB problem*, and, as we will see in the next subsections, is the most common framework assumed in RAR methods for CTs. In general, as MAB problems are a special class of MDPs, the traditional technique to address them is via a DP and backward induction. However, as shown in Section 3.1.1, such techniques suffer from a severe computational and memory burden, especially when size of the state space grows with the horizon T.

In order to provide the reader with an idea of existing RL and MAB-based strategies for response-adaptive randomization, we now illustrate some of the proposals in the field, focusing on the *Gittins index*'s approaches. We want to mention that common MAB techniques such as LinTS and LinUCB (see Section 3.1.2) have also been employed in RAR, and other new proposals also exists (see e.g., Atan *et al.*, 2019; Ahuja & Birge, 2016, 2020), also in different areas such as in two-stages SMART designs (Cheung *et al.*, 2015). However, unlike the comprehensive work for adaptive interventions which constituted one of the main focus of this thesis, the purpose of this section is more illustrative for setting the ground of the problem and related challenges which will be discussed in Chapter 4.

MAB approaches based on the Gittins index

The *Gittins index theorem* represents a cornerstone for most of the current work in adaptive CTs and a key breakthrough for the MAB problem in (3.30), as it showed how equivalently to solve it with k 1-dimensional MDPs instead of the k-dimensional MDP as required by DP. While the *Gittins index* (GI), originally named *dynamic allocation index* (Gittins, 1974), offers a solution to a very large number of problems (see Chapter 1 of Gittins *et al.*, 2011, for an overview), here we associate it with the Bayesian Bernoulli MAB problem illustrated above. More specifically, assuming that patients enter the trial one-by-one and the outcome for patient t is observed before patient $t+1$ appears, the GI theorem states that, for any infinite-horizon ($T = \infty$) discounted MAB problem with finitely many arms and bounded rewards, the allocation rule obtained by backward induction is optimal if and only if it always allocates patient $t+1$ to the treatment with the highest GI at time t. The GI of treatment k at state $x_{k,t}$, denoted by $\mathcal{G}_k(x_{k,t})$, is given by:

$$\mathcal{G}_k(x_{k,t}) = \sup_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{i=0}^{\tau-1} \gamma^i \mathbb{E}[p_k|x_{k,t+i}] \mid x_{k,t}\right]}{\mathbb{E}\left[\sum_{i=0}^{\tau-1} \gamma^i \mid x_{k,t}\right]}, \tag{3.31}$$

where τ is a $\{\sigma(X_{k,1}, \ldots, X_{k,t})\}_{t=1}^{\infty}$ (past-measurable) stopping-time, and the expectation is computed with respect to the Markovian states transition distribution (see Section 2.0.1). As illustrated above, states $x_{k,t} \in \mathcal{X}_{k,t} \doteq \{(s_{k,0} + s_{k,0}, f_{k,0} + F_{k,t} \in \mathbb{N}_+^2 : S_{k,t} + F_{k,t} \leq t, \forall t = 0, 1, \ldots, T\}$ represent the posterior distribution, i.e., all the possible two-dimensional vectors of information on the unknown parameter p_k

at time t: $(S_{k,t}, F_{k,t})$ is the random vector of successful and unsuccessful patient outcomes, and $(s_{k,0}, f_{k,0})$ is the prior hyperparameters.

The GI policy assigns a value to each treatment based on the observed state variables, and suggests as optimal strategy the one with the highest value. It can be calculated with off-line and on-line algorithms (see Chakravorty & Mahajan, 2014, for an overview), but typically it is computed by solving the problem of allocating patients optimally between treatment k and a known treatment which yields a constant reward (the detailed explanation is given in Gittins *et al.*, 2011).

Despite the optimality of the GI, the rule has never been used in real clinical trials. Practical barriers include: (1) the underline infinite-horizon setting; (2) insufficient statistical power (when computed using traditional hypothesis testing procedures); (3) the need for instantly observed treatment outcomes; and, (4) a lack of randomization to provide a basis for inference. We now review some of the subsequent literature that tried to overcome this limitations.

Restless MABs and the Whittle index. The Gittins index theorem was developed for a infinite-horizon *discounted Markov bandit problem*. However, in clinical trials we deal with finite resources or patients, typically, we choose the minimum number of patients to achieve a pre-determined power. Besides the requirement of infinite horizon, the theory of the GI only applies when the actions are deterministic. When these are randomized, the problem becomes so-called *restless*, meaning that more than one arm can change its state in every period. Also, even when the actions are deterministic, but the horizon is finite, the problem can be seen as restless, by adding the remaining number of subject allocations to the state of each arm. Based on this, an equivalent finite-horizon version of the GI for the Bayesian Bernoulli MAB problem was derived by reformulating the problem as an infinite-horizon *restless* MAB. A solution of the restless problem was given by Whittle *et al.* (1981), with the *Whittle index* (WI), which reduces to the GI in the non-restless setting.

Controlled Gittins. To overcome the severe loss of statistical power of the Gittins index, Villar *et al.* (2015a) introduced a controlled version of the GI, called *controlled Gittins* (CG) approach, that ensures that the allocation to the control treatment never goes below $1/(K+1)$. The procedure can be viewed as a composite design in which one in every $K+1$ patients is allocated to the control group, and the remaining patients are assigned to the experimental K treatments using the Gittins index rule. Based on simulation results, CG managed to solve the trade-off between power and patients benefits quite successfully, achieving more than 80% power with a mean number of successes higher than the one achieved by fixed randomization and other adaptive allocation strategies such as TS and less variability compared to TS.

The CG approach was initially developed for binary outcomes as in the GI, but later extended to normally distributed endpoints with known variance (Smith & Villar, 2018), based on the reformulation of the GI for continuous data (Gittins *et al.*, 2011).

Forward-looking Gittins index. Motivated by the deterministic and fully sequential nature of the Gittins index, which reduces its applicability to medical contexts where outcomes are observable soon after treating a patient, Villar *et al.* (2015b)

developed a probabilistic version of the GI, which can also be applied to blocks of patients rather than individual patients. Assume that a total of T patients are enrolled over J stages, each of which consisting in a block of size b so that $J \times b = T$. At each decision point j ($j = 1, \ldots, J$), the aim is to allocate the next b patients to the $K+1$ treatments, given the data up to block $j-1$; or equivalently to determine the probability of allocation to treatment k at stage j, $\pi_{k,j}$, which is common to all b patients of block j. Denoting with $\tilde{\boldsymbol{x}}_{(j-1)b}$ all the data observed up to block $j-1$, which can be written as a $(K+1) \times 2$ matrix in which row k represents the parameters of treatment k's current posterior distribution up to patient $(j-1)b$, adopting the GI, this probability if given by

$$\pi_{k,j} = \frac{1}{b} \sum_{t=(j-1)b+1}^{jb} \Big[\sum_{\tilde{\boldsymbol{x}}_{t-1} \in \tilde{\mathcal{X}}_{t-1}} \mathbb{P}(a_{k,t}^{\mathrm{GI}} = 1 | \tilde{\boldsymbol{X}}_{t-1} = \tilde{\boldsymbol{x}}_{t-1} \times \mathbb{P}(\tilde{\boldsymbol{X}}_{t-1} = \tilde{\boldsymbol{x}}_{t-1} | \tilde{\boldsymbol{X}}_{(j-1)b}) = \tilde{\boldsymbol{x}}_{(j-1)b}) \Big], \tag{3.32}$$

with $\tilde{\mathcal{X}}_{t-1}$ representing the set of all possible values for $\tilde{\boldsymbol{X}}_{t-1}$ given initial data $\tilde{\boldsymbol{x}}_{(j-1)b}$ for every future patient t in $(j-1)b+1, \ldots, jb$ under the GI rule (summarized by $a_{k,t}^{\mathrm{GI}}$). Note that the computational cost of computing the $\pi_{k,j}$'s, which depend on the joint state for the $K+1$ arms, i.e., $\tilde{\boldsymbol{x}}_t$ (instead of the one-arm state $\boldsymbol{x}_t$ as in the GI), will grow exponentially as b and K increase.

Randomization is introduced by taking into account future sequences of allocations in the same block under Gittins' rule; from here the name *forward-looking Gittins index* (LFGI). Indeed, notice that

- for all the b patients of block $j = 1$, $\tilde{x}_{(j-1)b}$ contains only the prior beliefs on the effectiveness of each treatment, and $\pi_{k,j} = 1/(K+1)$ for all k, if all treatments are assigned the same prior;
- for the first patient of block j ($j = 2, \ldots, J$), i.e., for patient $t = (j-1)b+1$), the allocation rule is deterministic if a single treatment has a unique maximum GI given $\tilde{x}_{(j-1)b}$;
- from the second patient of block j ($j = 2, \ldots, J$), i.e., for patients $t = (j-1)b+2, \ldots, jb$), allocation probabilities are obtained by averaging over the posterior predictive distribution (Beta-Bernoulli in case of Bernoulli data and Beta priors) of future data given $\tilde{x}_{(j-1)b}$.

If the maximum GI is not unique, and multiple treatments are joint maxima, ties are broken at random among the optimal treatments, introducing a degree of randomization also for the first patient of the subsequent blocks.

In simulation studies carried out in Villar *et al.* (2015b), compared to fixed randomization and other adaptive allocation rules such as TS, the FLGI, as well as the GI, showed an increased number of patients assigned to the better arm and an increased successes. On the other hand, FLGI and GI resulted also in a dramatically reduced power. However, when considering a controlled version of the FLGI analogous to the CG, which ensures that the allocation to the control treatment never goes below $1/(K+1)$, the power issue was corrected, sometimes also

improving upon the fixed randomization scheme. In terms of type-1 error, the FLGI approach showed to control it conservatively, with higher performances compared to the GI and TS.

An extension of the FLGI, originally developed for binary responses, to normally distributed outcomes, can be found in Williamson & Villar (2020). In addition to showing the benefits of the response-adaptive designs compared to traditional equal randomized design, authors also showed that there are efficiency and patient benefit gains of using allocation procedures with a continuous endpoint instead of a binary one. These gains persist even if an anticipated low rate of missing data due to deaths, dropouts, or complete responses is imputed online.

A major criticism of response-adaptive randomization is that, despite their ethical benefit, they are not suitable for performing reliable statistical inference in general. One of the most prominent recent arguments against their use is the concern that the Type-I error rate may not be controlled at the nominal level. This can be the case of a drift in patient characteristics over time independently of any treatment effect, and the traditional methods of analysis are used. An example of the first phenomenon is when the underlying prognosis of patients recruited in the early stages of a trial differs from those recruited in the latter stages. Using covariate-adjusted response-adaptive randomization can be a solution to this problem if the underlying covariates causing the heterogeneity are known in advance. For this reason, RAR is still infrequently used in practice. In the next chapter we fully illustrate and discuss this problem, and to conclude this section we illustrate one proposal on adaptive randomization based on time-to-event outcomes with covariates.

Continuous Bayesian adaptive randomization based on event times with covariates . Motivated by a practical problem that has arisen repeatedly when applying adaptive-randomization methods in trials when not all outcomes are observed immediately, Cheung *et al.* (2006) proposed a Thompson Sampling inspired approach with time-to event outcomes that also accounts for baseline prognostic covariates. Authors first assume the Weibull model, instead of the widely employed Bernoulli model, in which the survivor function for treatment k is denoted by $S_k(T|\mathbf{X};\boldsymbol{\theta}) \doteq \mathbb{P}(Y > T|\mathbf{X};\boldsymbol{\theta})$, with T a fixed time such that the patient's treatment is considered a success if $Y > T$, $\mathbf{X}$ the covariates and $\boldsymbol{\theta}$ the Weibull model parameters.

The AR criterion is based on the fact that treatment k is superior to treatment k' if $S_k(T|\mathbf{X};\boldsymbol{\theta}) \geq S_{k'}(T|\mathbf{X};\boldsymbol{\theta})$. Let $\mathbf{H_t}$ denote the data accumulated from patients in the trial up to study time t. If patient i with covariates $\mathbf{X}_i$ enters the trial at study time t_i, the AR criterion is defined as the posterior probability that treatment k is superior to all others, in terms of the probability of surviving beyond T given the covariate vector $\mathbf{X}_i$ of the patient accrued at t_i, i.e.,

$$\gamma_i(k|\mathbf{X}_i) = \mathbb{P}\Big(S_k(T|\mathbf{X}_i;\boldsymbol{\theta}) = \max_{j=1,\dots,K} S_j(T|\mathbf{X}_i;\boldsymbol{\theta})|\mathbf{H_{t_i}}\Big). \tag{3.33}$$

Since $\sum_{k=1}^{K} \gamma_i(k|\mathbf{X}_i) = 1$, when the survival time distributions are continuous, $\gamma_i(1|\mathbf{X}_i), \dots, \gamma_i(K|\mathbf{X}_i)$ may be used as the allocation probabilities. For $K = 2$ and no covariate, (3.33) is similar to the randomization probability proposed by Thompson (1933), who considered two independent binomial samples with probabilities

following beta priors. Here, a Weibull distribution is assumed for the time response variables Y, such that $\log\Big(-\log(S_k(y|\mathbf{X}_i;\boldsymbol{\theta})\Big) = \log(\mu_k) + \phi\log(y) + \boldsymbol{\beta}'\mathbf{X}$, where $\boldsymbol{\theta} = (\mu, \phi, \boldsymbol{\beta})$, with $\phi > 0$ and $\mu > 0$ the baseline rate parameter; and non-informative independent normal priors on $(\log(\mu); \log(\phi); \log(\boldsymbol{\beta}))$. In addition to this full model based approach, authors also illustrate the use of a semi-parametric approach (without assuming a distribution for Y and only requiring some mild assumption) based on *approximate Bayes methods*. We refer to the original work of Cheung *et al.* (2006) for readers interested in this proposal.

Chapter 4

Inference in Adaptively-Randomized Experiments with MABs[1]

Abstract

Multi-armed bandit algorithms have been argued for decades as useful for adaptively-randomized experiments. In such experiments, an algorithm varies which arms (e.g., alternative treatments or text-messages) are assigned to participants, with the goal of assigning higher-reward arms to as many participants as possible. However, existing works suggest that adaptive randomization leads to estimation bias and a poorer confidence intervals coverage, compared to uniform randomization.
By using a real-world 2-arm binary reward setting experiment as a motivating example, in this Chapter, we empirically investigate the impact of using Thompson Sampling (TS) instead of uniform random (UR) assignment. Focusing on hypothesis testing, with common test statistics for mean differences, we show that using TS can as much as double the type-I error (α; incorrectly reporting differences when none exist) and the type-II error (β; failing to report differences when they exist). We empirically illustrate how and why this occurs. Maximizing a reward function can lead to unbalance samples in favour to a random superior arm under the null, which creates type-I error inflation; and reduced confidence in estimates under the alternative, which inflates type-II error, thus reduces power ($1 - \beta$; correctly reporting differences when they exist). We show this problem persists using different bandit strategies and different test statistics, including a Bayesian framework. We then propose two adjusted versions of the Wald Z-test that incorporates knowledge of the adaptive randomization nature of the TS algorithm. While, these adjustments can help, they do not eliminate completely the inference issues. As an alternative strategy, we show that modifying the algorithm itself, by introducing a higher degree of uniform randomization when no empirical evidence for mean differences exists, may help in solving the issue. These results, primarily illustrate the nature of the problem, and then, suggest different strategies for improving statistical inference in

[1]Parts of the text of this chapter are extracted from the submitted/published manuscripts coauthored by the candidate and listed on page vii.

adaptive experiments based on multi-armed bandit algorithms.

4.1 Introduction

Multi-armed bandit algorithms have been put forth for decades as being useful for adaptively-randomized experiments, where an algorithm varies which arms (e.g., alternative interventions, treatments or text-messages) are assigned to participants, with the goal of giving higher-reward arms to as many participants as possible (Berry & Fristedt, 1985a). Furthermore, the rising usage of digital technologies enables scientists to conduct randomized experiments in real-world settings to understand people's real-world behavior and help them achieving their goals. For example, recent work has shown bandit algorithms can speed up use of data to help participants in education (Clement *et al.*, 2014, 2015; Williams *et al.*, 2016, 2018; Segal *et al.*, 2018), in healthcare (Tewari & Murphy, 2017; Rabbi *et al.*, 2015; Aguilera *et al.*, 2020), and in product design (Li *et al.*, 2010; Chapelle & Li, 2011; Lomas *et al.*, 2016). Yet, these examples are only a tiny fraction of the tens of thousands of experiments where bandit algorithms could be useful by directing more participants to more effective conditions.

In a multi-armed bandit problem, a system learns about the value of different arms by choosing among them, and receiving a stochastic reward associated with the chosen arm (see Section 2.0.1, or alternatively Lattimore & Szepesvári, 2020, for an overview). The mean reward for each arm is initially unknown (although rewards are independent of one another given the arm choices) and the system learns about the reward distributions based on its choices. Typical bandit algorithms aim to make arm choices that maximize the expected cumulative reward, or equivalently that minimize the cumulative regret, as reported in (2.13). Adaptively-randomized experiments (which we abbreviate sometimes to adaptive experiments) can be viewed as a bandit problem by considering the randomization of arms to participants. If the randomization is based on previous responses of participants, then, we have a response-adaptive randomization.

For example, in a clinical trial, arms might be the different available treatments, and the reward might be whether the patient responded to that treatment or not. To achieve the scientific goal of a randomized trial, e.g., of discovering whether one treatment is more effective compared to a control treatment, typically, patients are randomized to treatments uniformly at random (equal and fixed probability). To achieve the practical (and more ethical) goal of assigning the best treatment more often, an algorithm that dynamically modifies the randomization probability of future patients, by using the evidence of previous patients' responses, would be preferred. However, broader use of adaptive experiments, requires a better understanding of the trade-offs bandit algorithms make between the scientific and practical goal.

A major barrier to adopting bandit algorithms for experimental designs is the lack of clarity on how statistical analyses of data are impacted when using a bandit algorithm to adapt an experiment (Burnett *et al.*, 2020). Theoretical work suggests that adaptive data collection like the one used in bandit algorithms can induce bias in the estimates of means (Bowden & Trippa, 2017; Deshpande *et al.*, 2018; Nie *et al.*, 2018; Shin *et al.*, 2019, 2020), and that confidence intervals constructed

from these statistics may not have correct coverage (Hadad *et al.*, 2019; Zhang *et al.*, 2020b). Both practical decisions and scientific research rests on knowing and controlling how frequently type-I error occurs, as these can be deeply problematic. Incorrectly concluding one intervention is more effective than another can lead to wasted resources from a practical perspective, and mislead future research that builds on these findings. Recommendations for changes to practice typically rely on meta-analyses or multiple studies reporting similar findings (e.g., Means *et al.*, 2009), but an uncertainty about the inference performance has the potential to negatively impact the likelihood that research will be translated to practical improvements. It is therefore necessary to quantify how collecting data using a bandit algorithm (rather then a uniform randomization) impacts the type-I error of a statistical hypothesis test in a particular application, e.g., how often one incorrectly concludes the sample data provides evidence for a difference in arm means, when none exists.

Scientists or practitioners also need empirical insight into the impact on statistical power of a test, or the probability of concluding there is a difference in arm means when it truly exists. We remind that power is equal to $1 -$ type-II error (or $1 - \beta$), i.e., the probability of failing to conclude a difference exists, when it does. When compared arms have unequal sample sizes, because one arm was assigned to fewer participants, lower confidence in the sample estimate of the inferior arm mean can reduce statistical power for detecting differences (Villar *et al.*, 2015a; Zhang *et al.*, 2020b; Yao *et al.*, 2020). Quantifying this reduction is also essential, as there is a high bar for scientists and practitioners to trust statistical analysess of adaptively-collected data, to ensure best practices are followed and meet regulatory requirements (FDA, 2019). Integrating statistical considerations to lower the barriers for analysis of real-world experiments by scientists and practitioners has tremendous opportunity for bandit algorithms to have an impact.

This work therefore aims to provide empirical evaluation into how using bandit algorithms impacts statistical hypothesis testing. We explore in a simulations study how much and why using a bandit algorithm inflates type-I error and reduces statistical power. We target the 2-arm binary reward setting, because it is ubiquitous in experiments, and because if these issues are non-trivial to solve in this case, they will only be more compounded in more complex settings. We constructed a simulation environment with parameterization of arm differences and number of participants inspired by real-world experiments (Williams *et al.*, 2016).

We show that when there is no difference in arms (e.g., both arm means have success probability of 0.5), using Thompson Sampling instead of uniform random can increase type-I error from 5% to as much as 13%. When there is a difference in the arm means (e.g., arm mean of 0.45 vs 0.55), we show that the power of a test to detect this effect is reduced from 80% with uniform randominzation to 56% with Thompson Sampling. We also show that the identified problems occur: 1) not only for the commonly used Wald Z-test (Wald, 1943), but also for the Welch's t-test (Welch, 1947), and when using a Bayesian framework for hypothesis testing, i.e., the Bayes Factor; 2) when using another common bandit strategy, i.e., ϵ-Greedy. We illustrate the potential motivation behind this trend.

These findings provided guidance to two alternative ways of pursuing improved statistical inference in data collected by bandit algorithms. First, to explore ways of modifying the test statistic using knowledge of the data collection process of

the bandit algorithm. More specifically, this work investigates two methods for adjusting the Wald Z-test: 1) using *inverse probability weighting* to reduce bias in the estimates of the means in an attempt to deflate the type-I error and increase power; 2) estimating through simulations the empirical distribution of Wald Z-test, assuming data are collected by TS, and use the latter, rather then the Wald Z-test's theoretical normal distribution. Second, to modify the algorithm or framework for formulating the problem of interest: being more sensitive to statistical analysis as well as to reward maximization.

These analyses and results can provide insights into the challenges to be surmounted in bridging machine learning, statistics, and applied sciences, to conduct adaptive experiments in the real-world, in an aim to simultaneously help individuals and advance scientific research.

In summary, the contributions of this Chapter are:

- To empirically investigate the challenges of drawing inferences, specifically hypothesis testing, from bandit-collected data in simulation, and show that these issues persist across different hypothesis tests and bandit strategies;
- To give insights and explain why adaptively-collected data negatively impact type-I error and power of a statistical test;
- To explore two ways of modifying a common traditional statistical test, i.e., the Wald Z-test, by incorporating knowledge of the adaptive nature of bandit algorithms; and show that, while these can mitigate some of the arising issues, they do not fully solve the hypothesis testing problem;
- To propose a new bandit framework that balances reward maximization and uniform randomization.

Our hope is to illustrate some considerations necessary for applying bandit algorithms for both maximizing reward and enabling reliable statistical analysis and inference from data, as well as inspire the development and modification of theoretical frameworks and algorithms to better tackle these issues.

4.2 Related Work

Real-world applications of bandit algorithms. Bandit algorithms have been applied to conduct adaptive experiments in different areas where maximizing the immediate participants experience takes precedence over generalizable statistical conclusions from collected data. This is the classic bandit problem formulation of maximizing reward, or, equivalently, minimizing regret. Applications include both industry, e.g., to give more popular versions of websites (Hauser *et al.*, 2009; White, 2012), product features or advertisements (Li *et al.*, 2010; Chapelle & Li, 2011; Russo *et al.*, 2018; Bakshy *et al.*, 2012) or to find the best available radio channel from a large set of channels (Toldov *et al.*, 2016), and research, e.g., in education (Clement *et al.*, 2015; Segal *et al.*, 2018; Williams *et al.*, 2016). Here, the primary focus is on optimizing learning outcomes rather than questions of how best to analyze the

data from the experiments. However, it should be noticed that there are still many product teams that do not use bandit algorithms due to concerns about drawing generalizable conclusions from the data (Kohavi *et al.*, 2012).

Bandit algorithms, typically contextual bandits (see Chapter 2.0.1, or Tewari & Murphy, 2017, for an introduction), have also been applied in health research, such as to deliver text messages in order to improve users' physical activity (we present our related study in Chapter 5), or maximizing stress reduction (Paredes *et al.*, 2014). We don't focus on the rich literature on contextual bandit algorithms, but we expect that issues that arise in simpler cases are likely a compounded problem, probably with a higher impact, for more complex settings such as contextual bandit problems. These applications mainly belong in a bandit setting where optimization of reward is the key goal. However, a different line of literature also addressed the problem in the context of *best-arm identification*, e.g., for automate machine learning (Hoffman *et al.*, 2014), for selecting influenza mitigation strategies (Libin *et al.*, 2018) or for insect control for organic agriculture (Libin *et al.*, 2019), where the aim is to identify the arm with the highest mean with high confidence.

Despite the clear promise for using data more rapidly, the breadth of applications could be substantially increased if there was not the sense that one had to trade-off maximizing reward and drawing generalizable conclusions with reliably quantifiable levels of statistical certainty. Poor understanding of this aspect and absence of robust inference and estimation in adaptively collected data, constitutes one of the main drivers that prevents for instance the practical use of bandit strategies in clinical trials (Pallmann *et al.*, 2018; Burnett *et al.*, 2020). We now turn to related work on this topic and explain how the current work aims to advance this goal.

Statistical inference from bandit-collected data. In high stakes settings, such as clinical trials, there has been limited use of adaptive designs due to the challenges of drawing inferences from them (Pallmann *et al.*, 2018; Burnett *et al.*, 2020). While some theoretical and modeling work in this area exists (as shown in Chapter 3, Section 3.2), current guidance on adaptive clinical trials (FDA, 2019) emphasizes the need for more applications, case studies, and data sets on which to evaluate the theoretical work. Still, uptake of adaptive clinical designs is relatively slow, evidence that higher guarantees are needed for running trials, and lowering their risk of being inconclusive or wrongly conclusive (Burnett *et al.*, 2020). One barrier to the application of these methods is the ongoing debate and concerns about how such adaptive experiments influence properties of statistical hypothesis tests such as type-I error and statistical power. For a scientist, a lack of understanding of how much using a bandit algorithm impacts type-I error is deeply problematic, as they might draw an incorrect conclusion that a difference between arms or treatments exists when none occurs, which misleads future scientific research and contributes to the replication crisis (Camerer *et al.*, 2016; Collaboration *et al.*, 2015). Even practically, it can waste tremendous practical resources to pursue an intervention that seems to be better but is not actually better. Or, a low power where one misses an effective intervention.

Drawing inferences from bandit-collected data can be challenging due to biases in the means and other functions of the arm means (Atkinson *et al.*, 2014; Bowden & Trippa, 2017; Deshpande *et al.*, 2018; Nie *et al.*, 2018; Shin *et al.*, 2019, 2020). Thus,

some recent work, in the context of hypothesis testing, claims the need for unbiased estimators of means (e.g., Deshpande *et al.*, 2018; Hadad *et al.*, 2019; Zhang *et al.*, 2020b), which may also solve poor confidence intervals coverage (Hadad *et al.*, 2019; Zhang *et al.*, 2020b), rather than directly addressing type-I error and power. In this paper, we also explore this strategy, building on Bowden & Trippa (2017)'s work that uses inverse probability weighting (Robins, 2000) to reduce bias. However, this bias reduction inflates the variance of the estimator, which means there are still challenges in drawing conclusions about the relative values of different arms.

Some works have directly investigated the challenges in hypothesis testing with data from adaptive experiments using bandit algorithms (Villar *et al.*, 2015a; Zhang *et al.*, 2020b; Yao *et al.*, 2020; Kasy & Sautmann, 2021). Some of the proposed techniques rely on considerable information about (and control over) the data generating process. Yao *et al.* (2020) consider clipping the probabilities of selecting each action at each point in the data selection, and (Kasy & Sautmann, 2021) propose modifying the bandit algorithm so that an action is assigned at most half of the times to encourage more exploration. Other works, by focusing on addressing the biased ordinary least squared (OLS) estimator, require the computation of more advanced estimators, based on adaptive weights and other correlation summaries (Deshpande *et al.*, 2018; Hadad *et al.*, 2019; Zhang *et al.*, 2020b). In the spirit of this recent theoretical work, our primary goal is to provide empirical exploration and insights on the type-I error and power issue. Secondarily, we propose some alternative solutions, including a bandit modification strategy, to balance reward maximization and generalizability of conclusions. Our study considers various scenarios and uses different existing bandit algorithms and statistical tests. Our focus is complementary in exploring primarily the simplest form of TS (without parameters scientists would have to fit), and the widely used Wald Z-test, and targeting the 2-arm case with binary rewards.

Balancing reward maximization against other objectives. A theme of the current paper is exploring how reward-maximizing bandit algorithms impact the objective of statistical inference, and how this is similar to and different from past works that considered adaptive allocation to optimize other objectives besides regret. For instance, best-arm identification aims to adaptively assign arms in order to efficiently or accurately identify the optimal arm (Even-Dar *et al.*, 2002; Audibert & Bubeck, 2010; Russo *et al.*, 2018). Similarly, work on optimal experimental designs aims to estimate a parameter of interest with maximum precision and efficiency (Myung *et al.*, 2013; Kaptein, 2015; Smucker *et al.*, 2018). As algorithms cannot be guaranteed to be optimal on both reward maximization and other objectives, such as best-arm identification (Bubeck *et al.*, 2009), some literature has directly considered the question of how to trade off competing goals. For instance, some work proposed to use a multi-objective bandit problem for trading off cumulative reward against minimizing estimation errors for arm means rewards (Liu *et al.*, 2014; Erraqabi *et al.*, 2017) or to introduce a random cost for pulling an arm and constrain the total cost by a budget (Xia *et al.*, 2015; Hoffman *et al.*, 2014). Other work aims to maximize reward by choosing the best arm, while also gaining enough accuracy in estimating alternative arms to be able to have high confidence that the best arm was chosen and justify generalization about the best arm (Yang *et al.*, 2017; Jamieson & Jain, 2018).

In this paper, we also consider both reward and an additional objective, related to but different from the previous work. More specifically, our additional goal is to have low type-I error and high power when testing the null hypothesis of no difference between two arms. We focus specifically on understanding how an interpretable bandit algorithm (i.e., Thompson Sampling) performs on this alternative objective when adaptively collecting data. We believe that such empirical evaluations may give useful insights on how an hypothesis testing procedure is affected by the TS algorithm in different plausible settings and how to potentially correct it and ensure stronger guarantees. Existing literature provided only a partial view on this phenomenon, focusing mainly on modifying the bandit strategy in specific settings, which will represent our secondary goal in this work.

4.3 Challenges in Drawing Inferences from Data Collected with MABs

In this section, first, we introduce: 1) the compared allocation strategies, with a full illustration of the standard *Beta-Bernoulli Thompson Sampling* (a simplified version of Algorithm 4), designed to solve the MAB problem in a 2-arm binary reward setting; 2) the compared statistical procedures for the hypothesis testing problem, and; 3) the simulation environment. Then, we outline the harms to the Wald Z-test's type-I error and power that Beta-Bernoulli TS can cause when used to adaptively assign experiment participants to different conditions. We show that the problem persists when using different bandit strategies and different hypothesis testing procedure.

4.3.1 Methods and Simulation Environment

Allocation Strategies

We simulate our experiments by using two adaptive allocation strategies, i.e., Thompson Sampling and ϵ-Greedy, and compare them with a traditional uniform random allocation. We remind that a uniform random allocation will randomize participants to arms with equal probability, while the ϵ-Greedy algorithm will assign, with probability $1-\epsilon$, the arm with the highest mean reward so far (in the sample), and, with probability ϵ, a random arm. We consider a value of $\epsilon = 0.1$, so that for each participant, with probability 0.1 arms will be assigned using UR, and with probability 0.9 the greedy arm will be selected. For Thompson Sampling, given the binary reward and 2-arm setting, we assume a Beta-Bernoulli model and a $Beta(1,1)$ prior, corresponding to a uniform distribution. As TS will be the primary focus of the next section, we provide a detailed description of the Beta-Bernoulli Thompson Sampling.

Beta-Bernoulli Thompson Sampling. In 2-arm Beta-Bernoulli TS, pulling arm k, with $k \in \{1,2\}$, results in a reward of 1 with a probability p_k and 0 with probability $1-p_k$. p_k denotes the success rate of arm k, which is unknown and independent of the other arm. An independent beta-distributed prior with parameters of $\alpha_k > 0$ and $\beta_k > 0$ over the estimation of each p_k is assumed. At each iteration of TS, a

sample is drawn from the posterior distribution of p_k for each arm, and the arm with the larger sample is selected. This rule is equivalent to choosing arms with probability equal to the posterior probability that they have the highest probability of returning the highest reward (Chapelle & Li, 2011; Russo *et al.*, 2018), in our case a reward of 1. In other words, denoted with a_t and y_t the arm and the reward at time t, respectively, we have that the probability of selecting arm 1 at time t, say $\rho_{1,t} = \mathbb{P}(a_t = 1)$, is equivalent to its probability of being optimal, and in a 2-arm case can be defined as:

$$\begin{aligned}\rho_{1,t} &= \int \mathbb{I}\big[\mathbb{E}[y_t|a_t = 1, p_1] = \max_k \mathbb{E}[y_t|a_t = k, p_k]\big]\pi(\mathbf{p}|\mathcal{D}_t)d\mathbf{p} \\ &= \int_{[0,1]^2} \mathbb{I}\big[\mathbb{E}[y_t|a_t = 1, p_1] > \mathbb{E}[y_t|a_t = 2, p_2]\big]\pi(\mathbf{p}|\mathcal{D}_t)d\mathbf{p} \\ &= \int_{[0,1]^2} \mathbb{I}[p_1 > p_2]\pi(\mathbf{p}|\mathcal{D}_t)d\mathbf{p},\end{aligned}$$

where $\mathbb{I}$ is the indicator function, $\mathbf{p} = (p_1, p_2)$ is the parameters' vector, $\mathcal{D}_t$ is set of available data up to time t, and $\pi(\mathbf{p}|\mathcal{D}_t)$ is the posterior distribution of the unknown p_k's parameters given the observed data $\mathcal{D}_t$. Note that the second equality comes from the adaptation in a 2-arm setting, and the third one derives from the exact specification of the expected value of Binomial rewards.

After pulling arm a_t at time t and observing the associated reward y_t, the distribution of the selected arm is updated based on Bayes' rule, while the distribution for the other arm will stay the same, i.e,

$$(\alpha_k, \beta_k) \leftarrow \begin{cases}(\alpha_k, \beta_k) & \text{if } a_t \neq k \\ (\alpha_k, \beta_k) + (y_t, 1 - y_t) & \text{if } a_t = k.\end{cases} \tag{4.1}$$

Choosing actions with TS, balances exploration and exploitation in the long run, sampling from arms such that it converges on the better arm asymptotically (Agrawal & Goyal, 2012).

Hypothesis Testing Procedures

A typical example of interest in experimentation aims to test whether there is a difference between average outcomes across groups or not. The logic of hypothesis testing for a difference between two arm means, or two proportions, say p_1 and p_2, in case of binary rewards, consists of testing the null hypothesis $H_0 : p_1 - p_2 = 0$ of equal means against the alternative $H_1 : p_1 - p_2 \neq 0$ of different means. Scientists want to draw conclusions from a hypothesis test about whether there is evidence for a difference in the arms, or if a difference in sample of arm means results purely from chance.

Wald Z-test. One decision rule for rejecting the null hypothesis is by identifying a region of acceptance by using the Wald Z-test statistic (Wald, 1943), given by

$$Z = \frac{(\hat{p}_2 - \hat{p}_1)}{\sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}}, \tag{4.2}$$

where n_1 and n_2 are the number of the sample observations, $\hat{p}_1$ and $\hat{p}_2$ are the maximum likelihood estimators (MLEs), in the normal case the sample proportions (or means), of arms 1 and 2, respectively, which has a standard normal distribution under the null. Notice that in our case of binary data, we used the normal approximation of binomial distribution for n large enough (Peizer & Pratt, 1968). Such tests are thus computed through the comparison of the sample means and their respective number of participants - which also determines the standard error of the estimate for a proportion - and, given an acceptable significance level α, they give us a sense of the strength of the evidence that there is truly a difference. More specifically, we compute the observed test statistic, say z_{obs}, which is based on the observed arms and rewards data, and check if it belongs to the rejection or acceptance region, which is basically defined by the pre-specified α level. We use a significance level α of 0.05, as it is commonly used in social and behavioral sciences and was defined as a convenient cutoff level to reject the null hypothesis by Fisher (1992). We then reject the null hypothesis if $|z_{obs}| > F^{-1}(0.975) = 1.96$, where F^{-1} denotes the inverse Cumulative Density Function (CDF) of a standard normal distribution. The normal distribution is indeed the distribution of the Wald Z-test statistic under the null hypothesis, and assuming that the n_k's, with $k = \{1, 2\}$, are independent and identically distributed (i.i.d.).

We used as primary test statistic the Wald Z-test, motivated by the fact that it is a widely used test statistic and its asymptotic characteristics have been well investigated by statisticians, both in i.i.d. settings (Engle, 1984) and non i.i.d. data (Yi & Wang, 2011). In general, as sample sizes approach infinity, the Wald, Likelihood Ratio test, t-tests and Lagrange multiplier tests are equivalent; $n = 30$ gives already a good equivalence (Agresti, 2003). However, in adaptively collected data, Yi & Wang (2011) showed that the Wald Z-test has a better performance in statistical power for small to moderate sample sizes.

Welch's (unequal variances) t-test. Using the same strategy illustrated above, we now draw hypothesis testing conclusion by employing a different test statistic that is designed for unknown and unequal sample distribution variance (that may be caused by unbalances sample sizes), but requires the same assumption of sample distribution normality as in the Wald Z-test. The Welch's t-test (Welch, 1947) defines the test statistic as:

$$\frac{(\bar{p}_2 - \bar{p}_1)}{\sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}}},$$

where n_1 and n_2 are the sample sizes, $\bar{p}_1$ and $\bar{p}_2$ the sample means, and s_1 and s_2 the sample standard deviations, of arms 1 and 2, respectively. It represents an adaptions of the Student's t-test (the denominator is not based on a pooled variance estimate), and is more reliable, with higher control of the type-I error, when the two samples have unequal variances and/or unequal sample sizes (Ruxton, 2006; Derrick *et al.*, 2016).

Bayes Factor. As an alternative strategy we investigated a Bayesian framework for hypothesis testing. A Bayesian analysis compares two hypotheses as a special case of model comparison, providing a measure for which model (i.e., hypothesis)

better fits the data and quantifying the strength of that support (Rubin, 1978). We used the Bayes factor (BF) for comparing the evidence in favor of the alternative hypothesis that there is a difference in arm means with the null hypothesis that there is not a difference between arm means. The BF is given by the ratio of the marginal likelihoods for both hypothesis-specific models (Rubin, 1978). Details on the computation of the BF in the binary reward setting and the prior choices are given in Appendix D. We consider two cutoffs as the "critical" values for favouring the alternative hypothesis over the null hypothesis: 1) a threshold of 1, which is the more intuitive cut-point for selecting one of the two models and, 2) a threshold of 3. The choice of 3 is based on Jeffreys' scales of evidence for model selection (Jeffreys, 1961; Kass & Raftery, 1995), which considers a Bayes factor > 3 as substantial evidence in favour of the alternative hypothesis or the null. While the idea of evaluating type-I error and power with this approach results in a combination of Bayesian and frequentist analysis methods, rather than purely Bayesian, its results can help to illustrate that the issues with frequentist hypothesis testing in this setting are not caused only by idiosyncrasies of the tests we examined.

Simulation Environment

We focus on the kinds of data and analysis methods typically used in real-world experiments. For our 2-arm setting with a binary reward outcome, we construct a simulation environment that allows for varying the values and differences in arm means, as well as the sample size or number of trials. We examine cases where the arms have equal rates of rewards ($p_1 = p_2 = 0.5$) and where there is a difference of 0.1 in the reward rate ($p_1 = 0.55$, $p_2 = 0.45$). The latter corresponds to a small effect size, as measured by Cohen's w (Cohen, 1988); in many cases, effect sizes are quite small in experiments in social/behavioral sciences. In all simulations, we use a sample size $n = 785$ simulated participants. This is the sample size needed for uniform randomly collected data to have 80% power given the true arm differences that we use. We then conduct 5000 simulations for each assignment or data collection strategy (TS, UR, EG), and each arm difference (0, 0.1). This simulated the application of UR/TS/EG to 5000 randomized experiments with a particular experimental design ($n = 785$, 2 arms, binary reward) and two different arm differences, i.e., 0 and 0.1. In each simulated dataset, we compare the different allocations strategies and the different hypothesis testing procedures with respect to their type-I error and power. We compute type-I error as the proportion of the 5000 simulated experiments with true arm difference 0, where the statistical hypothesis test concluded there was a difference in arm means (e.g., value of the observed Wald Z-test statistic was ± 1.96 or greater in absolute value). Power is computed as the proportion of the 5000 simulations with true arm difference of 0.1, in which the statistical hypothesis test concluded there was a difference in arm means.

4.3.2 Results: Type-I Error and Power

Figure 4.1 shows the type-I error and statistical power using different statistical tests for data collected using TS, ϵ-Greedy and Uniform Random. We start by focusing on our main bandit algorithm of interest, Thompson Sampling. We can

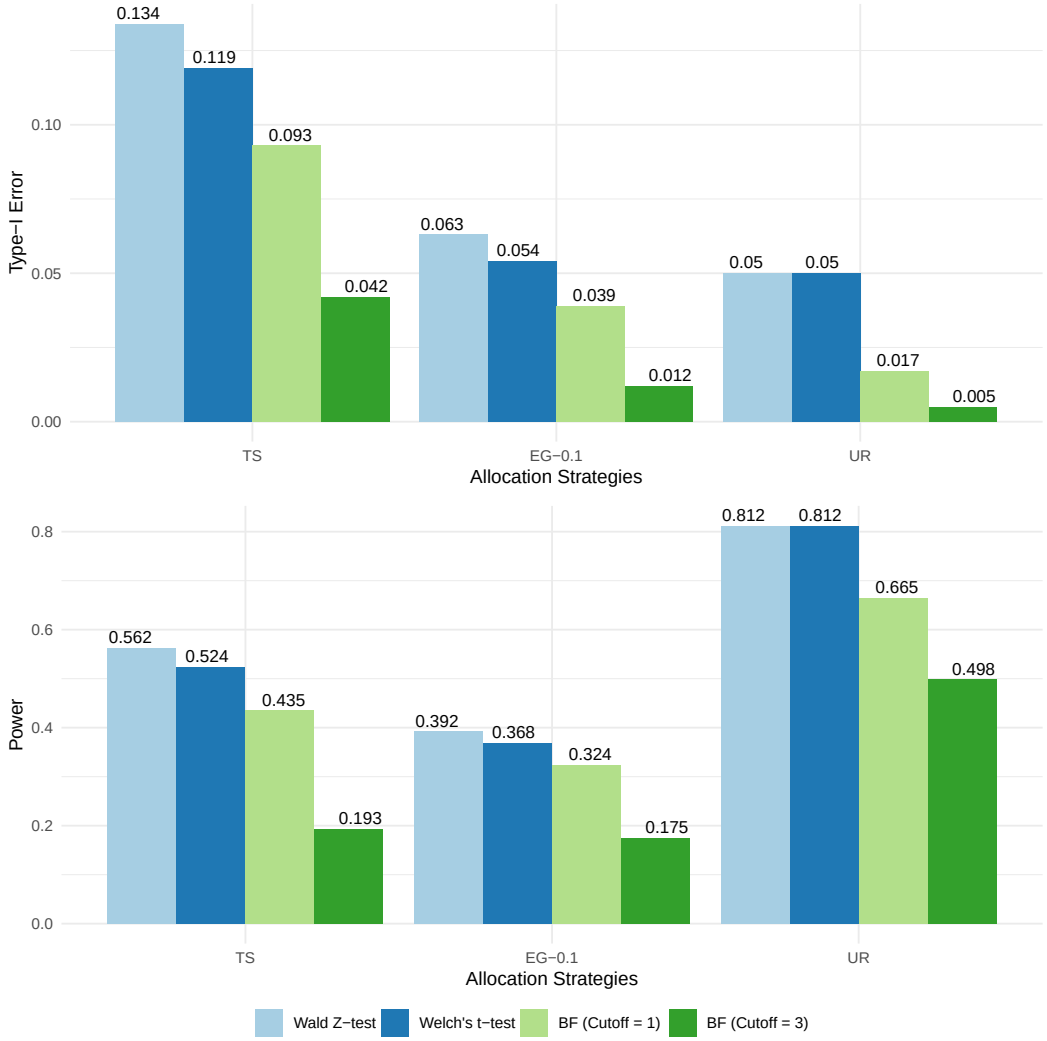

Figure 4.1. Type-I error and power for data collected using Thompson Sampling (TS), ϵ-Greedy with $\epsilon = 0.1$ (EG-0.1) and Uniform Random (UR). Hypothesis testing are based on a significance value $\alpha = 0.5$ and are performed with different statistical tests: 1. Wald Z-test, 2. Welch's t-test, 3. Bayes factor (BF) with cutoff 1, and 4. Bayes factor (BF) with cutoff 3. Results are based on a sample size of $n = 785$ and a number of independently simulated dataset of 5000.

see that type-I error of data collected with TS is generally inflated. When there is no difference in arm means, as expected, using the Wald Z-test, the type-I error is controlled at 5% for Uniform Random, but it goes up to 13% for TS. Statistical power is also reduced. With the Wald Z-test, we achieve a power of only 56% for TS compared to 80% for UR. This might make practitioners reluctant to use a bandit algorithm, as the chance that a statistical analysis will not be able to detect a difference when it exists (type-II error) goes from 20% to 44%. When investigating a different prior specification, i.e., Jeffreys' prior (Jeffreys, 1961), in the Thompson Sampling algorithm, no substantial differences were noticed, suggesting consistency of TS results across different priors; see Appendix E for more details.

Drivers of Inflated Type-I Error and Reduced Power

When there is no difference in arm means, type-I error is increased by biased underestimation of the "inferior" arm and increasing confidence in the "superior" arm. As discussed in the previous section, when the arm difference is 0 (in this work we assumed both arm means have success probability of 0.5), using Thompson Sampling instead of Uniform Random can increase type-I error from 5% to as much as 13% with the Wald Z-test, which is problematic for false discoveries. Figure 4.2 shows one of the simulations of TS for an arm-mean difference of 0 that illustrates what drives the rejection of the null hypothesis when it is true and why the overall type-I error of TS is inflated compared to UR.

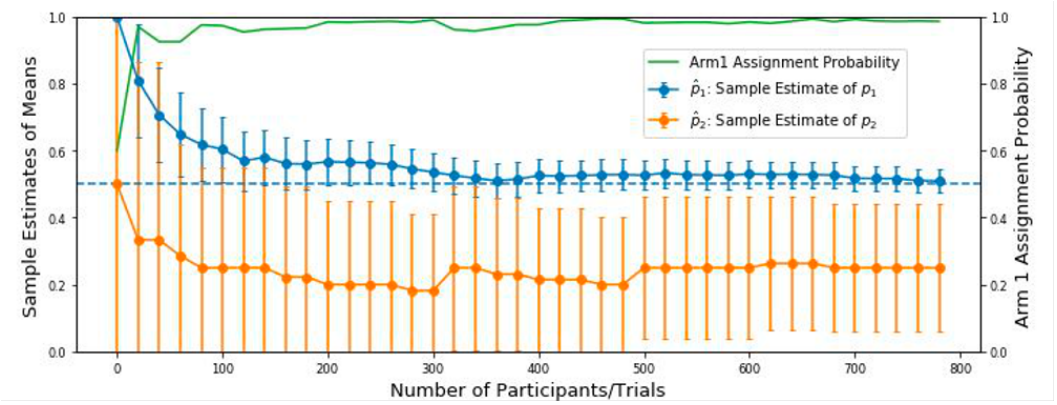

Figure 4.2. This figure shows an illustrative example of TS behaviour over the course of a single simulation experiment, which results in a type-I error. This is a scenario where the true arm difference is 0 with $p_1 = p_2 = 0.5$ and sample size $n = 785$. The sample mean estimates of arm 1 are displayed in blue, and the sample mean estimates of arm 2 are displayed in orange (left vertical axis). The assignment probability for arm 1 is denoted in green (right vertical axis). The vertical bars around the sample means represent 95% confidence intervals.

When there is no true underlying difference in arm means, sampling variability nevertheless will lead to sample estimates of means that make one arm look "inferior" (with a lower mean) than the other one. Since TS is adaptive to the posterior probability (which is sensitive to the sample mean), when (by chance) an arm's sample mean is lower than its true mean, TS will assign fewer participants to that arm, favouring the "superior" arm, which has a sample mean closer to (or higher than) its true mean. Increasingly assigning the seemingly "superior" arm to participants reduces the standard error in the estimate of its sample mean and increases confidence in its value, even though in reality it is not higher - the "inferior" arm simply has a sample mean that is below its true mean. Even as more and more data is collected, TS does not substantially correct the earlier biased estimate of the estimated lower mean. Figure 4.2 illustrates how the assignment probability only gets higher and higher as confidence in the higher mean increases, and this convergence means there is increasingly less assignment of the "inferior" arm, and so little opportunity to obtain data that contradicts earlier misleading observations. Notably, these results suggest that the convergence of a bandit algorithm to one arm is not in itself strong evidence that there is actually a difference in arm means. The

adaptation of a reward-maximizing algorithm to random lows and highs can result in a very high assignment probability (and far more participants) being assigned to one arm. TS is more likely than UR to result in one arm having a higher sample mean than the other, even when the true means are identical.

When there actually exists a difference between arm means, the power to detect this difference is reduced because unequal assignment leads to reduced confidence in the estimate of the lower sample mean. When there is a difference in the arm means (in this work we assumed arm means of 0.55 and 0.45), we showed that statistical power of the Wald Z-test to detect this effect was reduced from 80% with Uniform Random to 56% with Thompson Sampling (see Figure 4.1). This means that trying to improve the reward for participants in an experiment takes the type-II error from 20% to 44%, more than doubling how often a UR experiment might fail to report a difference, when it exists. As shown in Figure 4.3, the power decreases because TS assigns very few samples into the truly worse arm, decreasing the overall confidence that a true difference exists.

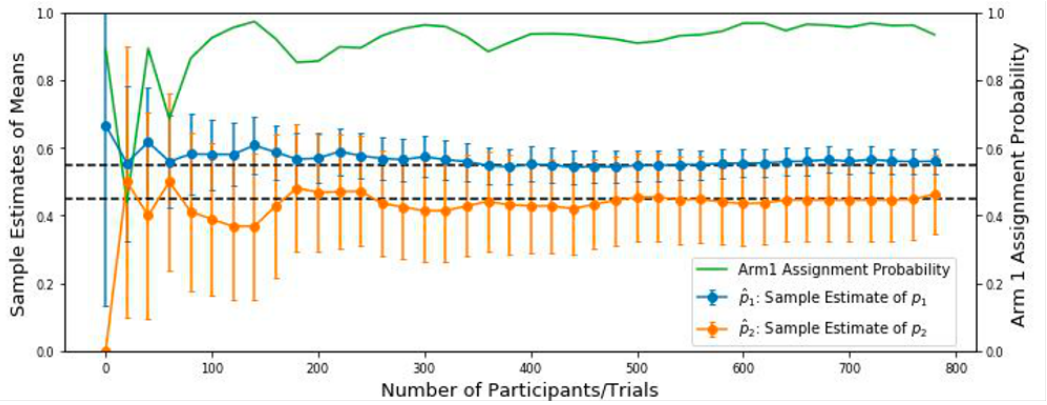

Figure 4.3. This figure shows an illustrative example of TS behaviour over the course of a single simulation, which results in a type-II error. This is a scenario where the true arm difference is 0.1 with $p_1 = 0.55$ and $p_2 = 0.45$ with sample size of $n = 785$. The sample mean estimates of arm 1 are displayed in blue, and sample mean estimates of arm 2 are displayed in orange (left vertical axis). The assignment probability for arm 1 is denoted in green (right vertical axis). The vertical bars around the sample means represent 95% confidence intervals.

Although TS underestimates the worse arm, which increases the estimated difference, this is outweighed by the far larger increase in the standard error (and confidence interval) for the sample mean. As TS allocates most of the participants into the superior arm, increasing the reliability of the estimate as shown by the confidence interval becoming tighter around the mean. However, great uncertainty persists in the sample mean for the inferior arm, illustrated by the confidence intervals remaining fairly large even after the experiment is complete. Since the inflation in the estimated difference is outweighed by the corresponding change in the standard error, this results in a smaller observed Wald Z-test statistic under the alternative, decreasing the power to reject the null hypothesis. Notably, the distribution of the Wald Z-test statistic, expected to be a standard gaussian, results in fatter tails under the null, and a shifted mean under the alternative (see Figure 4.4).

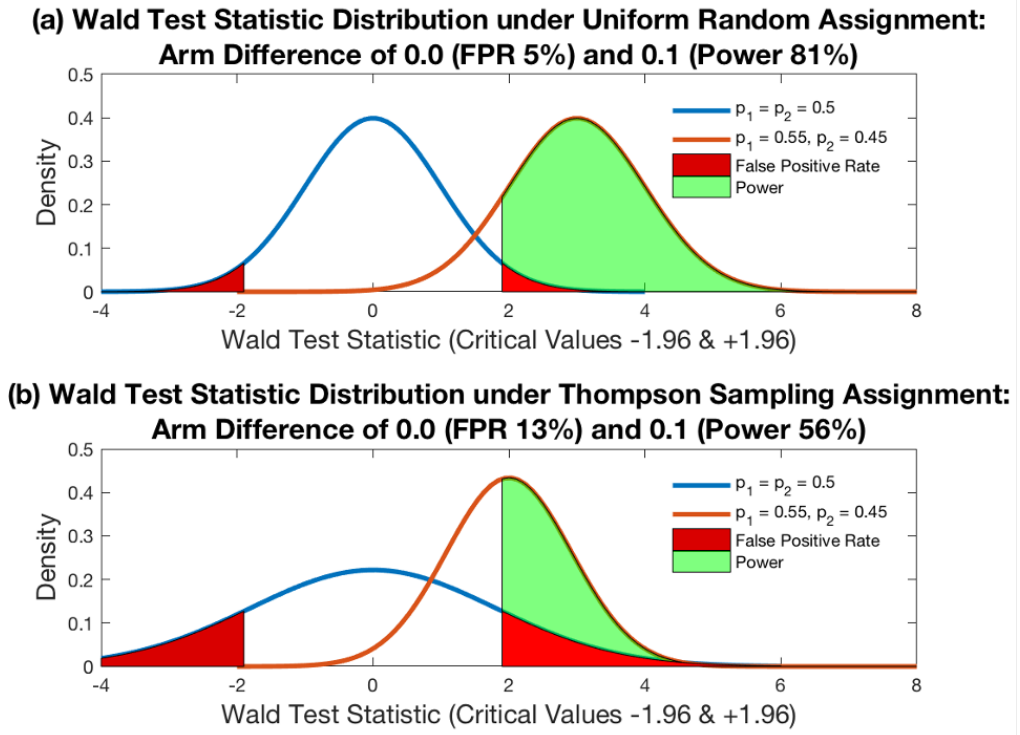

Figure 4.4. Distribution of the Wald test statistic under both the null and alternative hypothesis, for (a) Uniform Random (UR) allocation, and (b) Thompson Sampling (TS) allocation. The common critical value of 1.96 for hypothesis testing is used for the Wald Z-test statistic. Type-I error is shown as the red areas under the curve, and power as the green area. We can see that the type-I error is increased when using TS (compared to UR) because the distribution of the Wald Z-test statistic has fatter tails under arm difference of 0. Power is also decreased because the mean of the Wald Z-test statistic distribution is smaller under the alternative hypothesis compared to the Wald Z-test performed on UR data.

Higher assignment probability to one arm is not sufficiently indicative of an arm-means difference. One might hypothesize that when TS has a higher assignment probability for one arm and puts more participants in one arm, this higher posterior probability that the arm is optimal could in itself be more reliable evidence for a difference in arms, than the statistical hypothesis tests we have investigated. However, based on what we just discussed, we see evidence against this idea, and reason to see high assignment probabilities as providing less evidence for a difference than might be presumed. The results in Table 4.1 show that even when there is no difference in arm means, the assignment probabilities can mislead one to believe that an arm is superior.

For example, when the difference in arm means is 0, the assignment probability will be 0.7 or higher in the 70% of times. When the difference in arm means is 0.1, the assignment probability will be 0.7 or higher in the 98% of times. Table 4.1 shows what proportion of the 5000 simulations result in a particular assignment probability whichever arm has the higher assignment probability in a simulation. So a simulation where the superior arm had an assignment probability of 0.74 would contribute to

Assignment Probability of "Superior" Arm	Proportion of Simulations	
	Arm Difference of 0	Arm Difference of 0.1
$[0.5, 0.6]$	15%	1%
$(0.6, 0.7]$	15%	1%
$(0.7, 0.8]$	16%	3%
$(0.8, 0.9]$	20%	8%
$(0.9, 1.0]$	34%	87%

Table 4.1. Distribution of assignment probability (Assign. Prob.) into the "superior" estimated arm when the arm difference in 0 (column 2) and when there exists an arm difference of 0.1 (column 3). This table shows the proportion of simulations in which an arm has a certain assignment probability (column 1) at the completion of an experiment. To illustrate that assignment probabilities are still extreme when there is no difference in arm means, we bin by the "superior" arm (whichever arm has higher assignment probability in a simulation, regardless of what the true arm means are) at various assignment probability intervals.

the proportion of simulations in the interval $(0.7, 0.8]$. These results show that large disparities in the assignment probabilities across arms (e.g., arm 1 is 0.2 and arm 2 is 0.8) are weaker evidence for a difference actually existing, than might be hoped. Even when there is no difference in arm means, the assignment probabilities can get very high quite often: assignment probabilities of 0.9 or higher occur 33% of the time when there is no difference in arm means, as TS tends to converge on one arm. Therefore when the arm difference is 0, the posterior probabilities are unreliable for inferring whether a real difference across the arms exists due to adaptive policies being sensitive to random noise as a result of their reward maximizing nature.

We now incorporate in the discussion an alternative bandit strategy and alternative statistical tests and compare their performances in terms of type-I error and power.

A Comparison with ϵ-Greedy

The results presented so far suggest the relevant type-I error and power issues would arise for other bandit algorithms, because of how reward-maximization changes responsiveness to random highs and lows. Of course, the exact pattern of how much power is decreased and type-I error is increased could vary, and this is a key direction for future research. Here, we investigate the ϵ-Greedy algorithm, with $\epsilon = 0.1$ (EG-0.1), meaning that 10% of participants are assigned using UR, and the remaining 90% are given the greedy arm (the one with higher sample mean). One might predict EG-0.1 would perform well due to explicitly incorporating UR assignment. When using the Wald Z-test, Table 4.1 shows that EG-0.1 seems to have better type-I error compared to TS (6% vs 13%), which is also quite close to the 5% nominal level of UR. However, this comes at a cost of power, which is reduced to only 39%, compared to 56% for TS and 81% for UR. Because of the greedy assignment, far fewer participants are assigned to the "inferior" arm and the very small sample sizes, mean there is insufficient evidence for a statistical test to conclude there is a difference. In addition to EG-0.1's highly reduced power, EG-0.1 also fails to attenuate its adaptiveness in cases where there is less certainty about the difference, where for the sake of both reward and inference, it would be preferable

to do more exploration.

A Comparison with Alternative Hypothesis Testing Approaches

Analysis using Welch's t-test shows similar results as the Wald Z-test. While the prior results suggest that properties of the data itself, not specifics of the Wald Z-test, lead to the issues with type-I error and power, one might hypothesize that the issues could be resolved using a test better suited to experiments with unequal sample sizes, that may cause unequal variances. We therefore consider Welch's t-test, which is typically used for handling unequal variances and/or unequal sample sizes between groups, and so might be particularly suitable for an adaptive experimentation based on bandit algorithms. However, Welch's t-test does not correct the issues: as shown in Table 4.1, the type-I error is only 1.5% less than using the Wald Z-test, and this reduction comes with a reduction in power of 3.8%. This illustrates that the issue is with biases in estimates of means and reduced confidence in those estimates, as illustrated earlier. The Welch's t-test and Wald Z-test are asymptotically equivalent (and, as $n \to \infty$, they will reject the same cases). The advantage of Welch's t-test for handling unequal variances in the estimate of the sample mean is largely observed in small sample sizes, so we would not expect a huge difference in performances with a sample size similar to the one we have in experiments we typically do ($n = 785$, smallest is $n = 88$; see e.g., Williams *et al.*, 2016) as typically a heuristic is that the advantage of t-tests over z-tests rapidly diminish from a sample size of $n = 30$ and larger (Hogg *et al.*, 1977; Casella & Berger, 2002).

Bayesian analysis shows similar patterns to frequentist analysis. While both frequentist tests exhibit the same issues with type-I error and power, one might hypothesize that these issues would be eliminated or mitigated by using a Bayesian framework, as illustrated in Section 4.3.1. When using the commonly used Bayes factor (BF) for Bayesian hypothesis testing, as shown in Figure 4.1, we see that the test is overall significantly more conservative than the Wald Z-test even for UR sampling: using the cutoff of 3 gives a type-I error of only 0.5% and power of 49.8%. Relative to these values, TS sampling increases the type-I error to 4.2%, while further decreasing the power to 19.3%. The looser cutoff of 1, where 1 indicates equal support for both models is somewhat less conservative, shows the same trends of inflated type-I error and decreased power for TS. This demonstrates that pattern of results when evaluating the data using a Bayesian-inspired approach mirrors the pattern of the purely frequentist methods.

As a final remark in this section, we want to mention that an important issue when changing the hypothesis testing framework using the Bayes factor is that it is very high stakes in the scientific world, for example the *Food and Drug Administration* has strict requirements for reporting results for trials they fund (FDA, 2019), and issues like an inflated type-I error could be seen as serious risk for funding and publication. For a behavioral scientist, the recommendation to "simply try a different statistical technique" that neither they nor many others may closely understand, without direct empirical evidence and many other papers published using the technique, is a nonstarter.

4.4 Proposals for Improving Hypothesis Testing

Section 4.3 highlighted the problem of increased type-I error and decreased power in applying the Wald Z-test and other hypothesis testing procedures to data adaptively-collected via Thompson Sampling. To lower the barriers for bandit algorithms to be applied in real-world adaptive experiments by scientists, we consider: *How can we use features of the algorithm used to run an adaptive experiment to improve the statistical inference procedures?*, or, alternatively, *How can we modify features of the algorithm used to run an adaptive experiment to improve statistical inference from the data collected?*. We try to explore these questions in the next sections.

4.4.1 Adjusting Existing Statistical Tests

We start by focusing on the first question and examine two approaches which are related to the Wald Z-test statistic itself, to address the inflated type-I error and the reduced power problem. These approaches are similar in the sense that they both adjust the Wald Z-test statistic by incorporating knowledge about the data generating process that is induced by the adaptive nature of the bandit algorithm. However, while the first proposal relies on replacing the biased MLE used in the Wald Z-test equation in (4.2), the second proposal directly estimates the distribution of the test statistic in order to derive some adjusted critical values, based on which inference can be performed.

Most of the statistical tests, including the parametric Wald Z-test, are based on some underlying assumptions of the data, such as the i.i.d. assumption. In a traditional UR experiment, where data are i.i.d., the MLE has strong theoretical properties such as unbiasedness and normality distribution. However, in adaptive experiments, an extensive number of studies have demonstrated both that the unbiasedness property does not hold anymore (Bowden & Trippa, 2017; Deshpande *et al.*, 2018; Nie *et al.*, 2018; Shin *et al.*, 2019, 2020), proposing also some strategies to correct this bias, both that the derived test statistics does not satisfy the theoretical standard normal distribution (Jamieson & Jain, 2018; Hadad *et al.*, 2019; Zhang *et al.*, 2020b). Thus, by specifically addressing the two issues, we investigate two alternatives, which we call *IPW-adjusted Wald Z-test* and *TS-induced Wald Z-test* (this idea builds on a previous work of Smith & Villar, 2018), and show how they perform in an hypothesis testing problem.

IPW-adjusted Wald Z-test

With the IPW-adjusted Wald Z-test, we propose to replace the biased MLE in the Wald Z-test equation in (4.2) with an unbiased estimator in order to evaluate whether correcting the bias might also improve the type-I error and power. The unbiased estimator we use here is the *Inverse Probability of Weighting* (IPW; Robins, 2000) estimator, proposed first, in causal inference literature (Robins, 2000; Robins *et al.*, 1994), and then, adopted by (Bowden & Trippa, 2017) in the context of adaptive clinical trials with data collected by *Play-the-winner* (PW; Zelen, 1969; Robbins, 1952) strategy. We introduce IPW as a second way (compared to the MLE) of estimating the sample proportions, by incorporating knowledge of the assignment

probabilities of Thompson Sampling. More formally, in the context of an adaptive randomized experiment, if we now denote with n the total number of participants and with y_i the binary outcome observed for the i-th participant, the IPW estimator of the mean reward for arm k is given by:

$$\hat{p}_k^{\text{IPW}} = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i \frac{\delta_{ik}}{\pi_{ik}}}{\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{ik}}{\pi_{ik}}},$$

where π_{ik} is the randomization probability for student i to arm k, with $k = \{1, 2\}$, and δ_{ik} is the delta function, which takes value 1 if student i is assigned to version k, and 0 otherwise.

Intuitively, the IPW estimator for arm k is a weighted average of observed rewards from arm k, i.e., $\delta_{ik} y_i$, with weights given by the inverse of the probability for that participant to be assigned to condition k, i.e., $1/\pi_{ik}$: the higher the probability of receiving a specific treatment, the lower the weight of the reward. This allows a fairer comparison across arms with different sample sizes. By replacing now the biased MLE in the Wald Z-test formula with the unbiased IPW estimator we obtain the proposed IPW-adjusted Wald test, given below:

$$Z^{\text{IPW}} = \frac{(\hat{p}_2^{\text{IPW}} - \hat{p}_2^{\text{IPW}})}{\sqrt{\frac{\hat{p}_2^{\text{IPW}}(1-\hat{p}_2^{\text{IPW}})}{n_2} + \frac{\hat{p}_1^{\text{IPW}}(1-\hat{p}_1^{\text{IPW}})}{n_1}}},$$

with n_1 and n_2 denoting the sample size of arm 1 and 2, respectively.

Performance of the IPW-adjusted Wald Z-test. First, by comparing the IPW estimator with the MLE estimator, we can gain insights into whether and how the bias in the estimates of arm means in adaptive experiments, impacts statistical power and type-I error. As shown in Table 4.2, using the IPW estimator instead of the MLE helps us to correct the bias of the arms mean when data are adaptively collected. In particular, when there is no difference across arms and $n = 785$, we see that the bias in the estimate of arm 1 is only -0.0032 for the IPW estimator, compared to -0.0231 for the MLE estimator. This reduction is also verified when there is a difference between arm means (arm difference of 0.1; $p_1 = 0.55$, $p_2 = 0.45$; $n = 785$), where the bias of the IPW estimates for each arm are -0.0005 and -0.0079 respectively, compared to -0.0031 and -0.1452 respectively for the MLE estimates.

Given this reduction in bias, one might expect that we would then see an improvement in the type-I error. However, as one can see in Figure 4.5, IPW only slightly decreased type-I error. We do see a small reduction in type-I error for IPW relative to MLE (10% (SE = 0.004) vs 13% (SE = 0.005)), although this 2.4% reduction in type-I error does not get to the level a scientist has set the test to, and it comes at the cost of a 20% reduction in power (from 56% to 36%). In addition, this reduction is driven by a different pattern, rather than the bias reduction. Indeed, by looking at the bias in arm means difference, we can see that the same value of -0.0015 is achieved with both MLE and IPW estimator under TS assignment, and this is the parameter we are interested in; in addition it also builds the Wald Z-test (as its numerator is given by the arm means difference). The main difference is related

	Arm Difference of 0			Arm Difference of 0.1		
Estimate of	$\mathbf{MLE_{UR}}$	$\mathbf{MLE_{TS}}$	$\mathbf{IPW_{TS}}$	$\mathbf{MLE_{UR}}$	$\mathbf{MLE_{TS}}$	$\mathbf{IPW_{TS}}$
$p_1 - p_2$	-0.0003	-0.0015	-0.0015	0.0997	0.1421	0.1074
$\lvert p_1 - p_2 \rvert$	0.0279	0.0623	0.0015	0.0998	0.1445	0.1074
bias of p_1	0.0000	-0.0231	-0.0032	0.0002	-0.0031	-0.0005
bias of p_2	-0.0003	-0.0215	-0.0017	0.0005	-0.1452	-0.0079
SE of p_1	0.0003	0.0009	0.0009	0.0003	0.0004	0.0004
SE of p_2	0.0003	0.0009	0.0009	0.0003	0.0013	0.0019
SE of the Wald Z-test	0.0139	0.0395	0.0428	0.0141	0.0455	0.0583

Table 4.2. Estimated arm-means differences and absolute arm-means differences, with their bias and standard errors (SE), and the SE of the distribution of the Wald Z-test statistic for: 1. the Maximum Likelihood Estimator (MLE) - based both on the Uniform Random ($\mathbf{MLE_{UR}}$) and Thompson Sampling ($\mathbf{MLE_{TS}}$) assignment - and 2. the Inverse Probability Weighted (IPW) estimator based on the Thompson Sampling assignment ($\mathbf{IPW_{TS}}$). We report these results for both an arm difference of 0 and of 0.1. Estimates, bias and SE are computed based on 5000 simulated trajectories of size $n = 785$.

to the overall distribution and variability of the two test statistics, particularly on the tails. The mass probability we have on the tails, as shown in Figure 4.4 for the Wald Z-test distribution of a UR and TS with the MLE estimator, corresponds to the type-I error (considering the critical values of ± 1.96). With the IPW-adjusted Wald Z-test, this mass is lower than the one we have for the standard Wald Z-test with TS assignment, but still higher that the 5% probability we would have had with the standard Wald Z-test in a UR experiment. Using IPW slightly reduces type-I error (and not fully to the expected 5%) at the cost of decreased power: relative to the standard Wald Z-test with the MLE, power decreases from 56% to only 36% for the IPW-adjusted Wald Z-test (see Figure 4.5).

The decrease in power may also be understood based on the distribution of the two test statistics: despite the decreased bias in the estimate of the arm means difference, the increase in the overall variability of the estimates (see Table 4.2), which is translated into an increase standard error of the IPW-adjusted Wald Z-test compared to the MLE-based Wald Z-test (standard error of 0.058 vs 0.046, respectively; Table 4.2).

TS-induced Wald Z-test

The second statistical test adjustment alternative we propose is this work, is the TS-induced Wald Z-test. Instead of tackling the biased estimator, we now aim to tackle the theoretical asymptotic distribution of the Wald Z-test distribution (under standard assumptions on the data). Indeed, this is known to be a standard normal distribution under the null, with i.i.d. data. However, when this assumption does not hold, using the theoretical standard normal distribution for testing an hypothesis may lead to wrong conclusions. As previously illustrated in Figure 4.4, the empirical distribution is different form a standard normal under TS-collected data, with a higher variability compared to the standard normal distribution, probably because of the correlation in the reward variables induced by the adaptive algorithm. We thus propose a more flexible non-parametric approach: simulating and estimating the

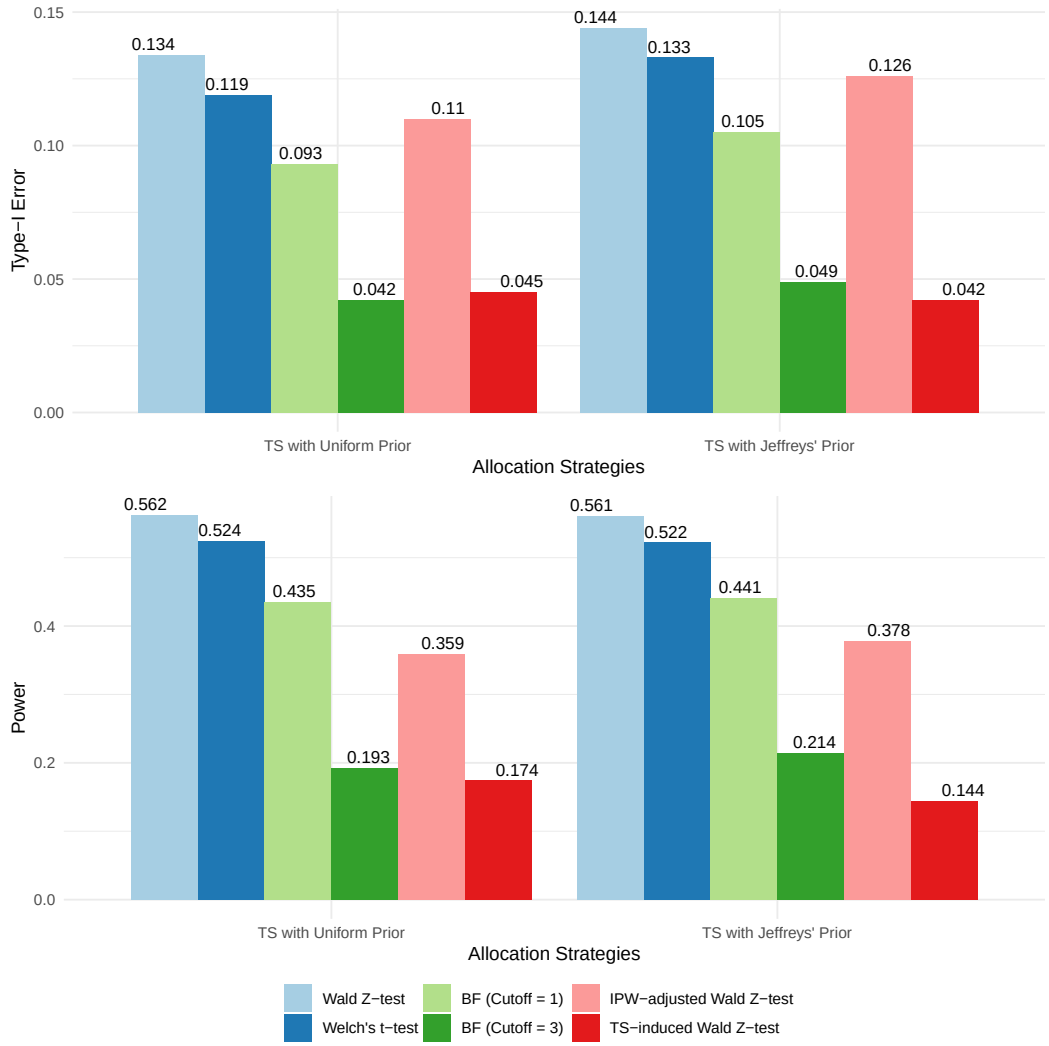

Figure 4.5. Type-I error and power for data collected using Thompson Sampling (TS) with two different prior choices, i.e., Beta$(1, 1)$, equivalent to a Uniform distribution in $[0, 1]$, and the Jeffreys' prior, in this case a Beta$(1/2, 1/2)$. Hypothesis testing are based on a significance value $\alpha = 0.5$ and are performed with different existing strategies, i.e., 1. Wald Z-test, 2. Welch's t-test, 3. Bayes factor (BF) with cutoff 1, 4. Bayes factor (BF) with cutoff 3, and the alternative strategies we propose, in this section, i.e., 5. IPW-adjusted Wald Z-test and 6. TS-induced Wald Z-test. Results are based on a sample size of $n = 785$ and a number of independently simulated dataset of 5000.

empirical distribution of the Wald Z-test under the null (when there is no difference in arm means; assuming $p_1 = p_2 = 0.5$ and $n = 785$) induced by the TS assignment procedure. We then derive its adjusted critical values as the empirical quantiles of the estimated TS-induced Wald Z-test distribution at the desired significance levels of $\alpha/2 = 0.025$ and $1 - \alpha/2 = 0.925$ for obtaining a 5% type-I error control.

As shown in Figure 4.5, the use of the TS-induced Wald Z-test allows type-I error to be controlled; compared to the standard Wald-Z test it is now reduced from 13% to 5%. However, as we can also see that this comes with a severe cost to statistical power, being only 17% for the proposed TS-induced Wald Z-test vs 56%

for the standard Wald Z-test with data collected adaptively with TS and 81% for the standard Wald Z-test with data collected with UR.

For checking the robustness of the proposed strategy we also change the simulation setting, based on which TS-induced Wald Z-test distribution is simulated and estimated. More specifically, we set the parameters $p_1 = p_2 = 0.25$ with the same sample size of $n = 785$, and replicate the above analysis. Appendix F shows that this has only a minor impact.

4.4.2 Adjusting the MAB Strategy: TS-PostDiff

In previous Section 4.4, we showed that both IPW-adjusted Wald Z-test and the TS-induced Wald Z-test can be useful in reducing the negative impact of a bandit algorithm on type-I error. We provided empirical illustration that can guide others to take this approach with a range of bandit algorithms and statistical tests. These methods for changing analysis techniques are attractive because they do not require changes to the algorithms' behavior, but simply employ knowledge of how the algorithm behaved. They do not impact participants' experiences, and can be employed anytime after the data are collected. However, the results also showed the statistical analysis problem cannot be easily or readily solved by simply modifying these tests. This suggested the importance of future work that modifies bandit algorithms (or the theoretical frameworks around maximizing reward) to change the data collection strategy, to be more sensitive to statistical considerations like type-I error and power, and the importance of considering whether different strategies are optimal based on the magnitude/existence of a difference in arms. We now discuss more in depth our proposal of formalizing the problem with a novel framework formulation, and propose a solution based on a modification of the Thompson Sampling algorithm we have discussed so far, which allows us to perform more reliable inference, particularly, in cases when there's no or very negligible difference between arms' means.

The novel problem of balancing reward maximization and inference

As discussed in Section 4.3.2, when there is small or no difference between arms, Thompson Sampling will often converge to assigning all participants to the same arm due to the variability of arms. In hypothesis testing, with traditional statistical tests such as the Wald Z-test, this behaviour can increase the probability of incorrectly rejecting the null hypothesis of no difference between arm means, compared to uniform random allocation. It can also reduce the probability of correctly rejecting the null hypothesis, which is especially impactful when effect sizes are small.

While it may seem intuitive to focus on large effect sizes, this may distract from the fact that small differences in arms can be extremely important to detect and use. For example, obviously large effects may be well known and easy to detect even without experimentation, while experimentation is necessary to discover small effect sizes. Even a 1% increase in the real-world can be extremely impactful. A great majority of effect sizes in many fields, such as education, are small. Nonetheless, discovering small effects can is still be scientifically valuable (Prentice & Miller, 1992). In fact, some have argued that seeking unrealistically large effect sizes has led

to poor statistical practice and lack of replicability (Fraley & Vazire, 2014; Friston, 2012). We would thus like an algorithm which is sensitive to the scenarios which are likely to arise in the real world, and can adjust its policy to increase the probability of distinguishing between a small effect and no effect.

More precisely, if an effect is small, the difference in reward between arms is not great, but the demand on statistical power in detecting this effect is large, and thus we would prefer the algorithm to assign participants to conditions closer to uniformly, to ensure we can detect this small effect. Likewise, when no effect exists, there is no reward to be gained, and reward maximization can result in an inflated type-I error. We would thus want to choose actions uniformly at random in this case. Finally, as the effect size grows larger, we would like the algorithm to assign more participants to the reward-maximizing arm. We thus introduce the novel problem of balancing the goal of a traditional randomized experiment for result generalizations (by favouring uniform random exploration when there's likely no chance to gain reward) and the MAB algorithms problem of reward maximization (by favouring exploitation, and assign more often the superior arm). Solving this problem is crucial to effectively designing adaptive experiments that also enable rigorous hypothesis testing with low type-I error and high power.

Proposed Strategy: TS with Posterior Difference Exploration

Motivated by the above, we would like to tune TS in order to increase uniform random assignment when there are non-existent or smaller effects (and less evidence for these effects), with a minimal loss in terms of reward when there is a substantial effect and reward to be gained. Towards this end, we propose the *Thompson Sampling with Posterior Difference Exploration* (TS-PostDiff) algorithm, which operates as follows:

- with probability ϕ_t, chooses arms with a UR policy
- with probability $1 - \phi_t$ chooses arms according to TS policy

We define ϕ_t to be the posterior probability after t steps that the difference in expected reward between actions is less than some threshold c. We propose to compute it as follows:

$$\begin{aligned}\phi_t &\doteq \mathbb{P}(|p_1 - p_2| < c|\mathcal{D}_t) \\ &= \int_{[0,1]^2} \mathbb{I}\big[|\mathbb{E}(y_t|a_t = 1, p_1) - \mathbb{E}(y_t|a_t = 2, p_2)| < c\big]\pi(\mathbf{p}|\mathcal{D}_t)\ d\mathbf{p} \\ &= \int_{[0,1]^2} \mathbb{I}\big[|p_1 - p_2| < c\big]\pi(\mathbf{p}|\mathcal{D}_t)\ d\mathbf{p},\end{aligned} \tag{4.3}$$

with $c \in (0, 1)$ a constant value set by the experimenter as a hyperparameter (interpretation of c is discussed below). As with TS, we can avoid computing (4.3) explicitly, and instead draw samples from our posterior distributions over rewards for each arm. However, note that in this case the focus is on computing the probability ϕ_t in a similar Bayesian fashion, rather that applying TS. This procedure results in choosing actions uniformly randomly if $|p_1 - p_2| < c$, and use TS otherwise; at step

Algorithm 7: TS-PostDiff

Input: $c \in (0, 1)$, $\alpha_k > 0$, $\beta_k > 0$, for $k \in \{1, 2\}$
for $t = 0, 1, 2, \ldots T$ **do**
  for $k \in \{1, 2\}$ **do**
    Sample $p_k \sim \text{Beta}(\alpha_k, \beta_k)$
  end for
  If $|p_1 - p_2| < c$ **then**
    $\tilde{a}_t = u + 1$, with $u \sim \text{Bern}(1/2)$
  else
  for $k \in \{1, 2\}$ **do**
    (Re-)sample $p_k \sim \text{Beta}(\alpha_k, \beta_k)$
  end for
  Select arm $\tilde{a}_t = \arg\max_k p_k$ and get the associated reward
  Update posterior parameters (α_k, β_k) for arm $\tilde{a}_t$ according to (4.1).
end for

t with probability ϕ_t a UR sceme is applied. Pseudo-code for TS-PostDiff is given in Algorithm 7.

Note the inclusion of a resampling step. This step is used to prevent using the same p_1, p_2 samples to determine if a difference exceeds c, as well as for arm selection, as this can result in TS behaving too exploitatively, choosing the estimated best arm too frequently.

Interpretation of hyperparameter c. Setting the hyperparameter c for TS-PostDiff allows us to constrain TS with respect to established small effect size thresholds in behavioural science. For instance, in behavioural science we find established thresholds for a small effect size: a Cohen's w of 0.1 is regarded as a small effect size (Cohen, 1988). We can thus set c to be around 0.1, which results in increasing power in the presence of small effects, increasing reward when effects are larger than c, and reducing type-I error inflation when no effect exists. We can think of c as the effect size below which we are willing to forgo reward in favour of improved statistical hypothesis testing.

Results

Under the same experiment setup as the one in Section 4.3.1 (arm differences of 0, with $p_1 = p_2 = 0.5$, and 0.1, with $p_1 = 0.55$ and $p_2 = 0.45$; $n = 785$), we now show in Table 4.3 the results comparing the standard TS, the proposed TS-PostDiff strategy with two alternative threshold values for the c parameter (0.1 and 0.2), and the UR gold standard procedure. In addition to reporting our quantities of interest, i.e., type-I error and power, we also look at the average reward and the proportion of optimal arm allocation.

Based on results in Table 4.3, we see that TS achieves the highest reward when there exist a small effect (0.536), but also the highest type-I error when there is no effect (0.135) and the lowest power when there is an effect (0.564). Also, we see that UR achieves controls the type-I error, achieving the highest power when there is an effect (0.806), but it is also associated with the lowest reward when there is an effect (0.50). Looking now at the values obtained by the TS-PostDiff strategy,

	Arm difference of 0	**Arm difference of** 0.1		
Method	Type-I Error	Power	Prop. Opt	Reward
UR	0.055 (0.002)	0.806 (0.004)	0.500 (0.0)	0.500 (0.0)
TS	0.135 (0.003)	0.564 (0.005)	0.860 (0.003)	0.536 (0.0)
TS-PostDiff ($c = 0.1$)	0.078 (0.003)	0.775 (0.004)	0.738 (0.003)	0.524 (0.0)
TS-Postdiff ($c = 0.2$)	0.054 (0.002)	0.800 (0.004)	0.560 (0.002)	0.506 (0.0)

Table 4.3. Table comparing UR (Uniform Random), TS (Thompson Sampling), and TS-PostDiff with exploration parameter c set to 0.1 and 0.2. Results for effect size 0 are shown in the left block (meaning actions have the same expected reward), and results for effect size 0.1 are shown in the right block. Within the 0 effect size block, we show type-I error, while in the effect size 0.1 block, we show power, mean reward, and the proportion of optimal allocation (Prop. Opt.) for each of the algorithms. In brackets one half the 95% confidence interval is reported. All shown values are averaged over 10,000 simulations.

we see that it achieves an effective compromise between the extremes of TS and UR for both considered threshold's c values. Notably, with $c = 0.1$, type-I error is below that of TS (0.078 vs 0.135), and little reward is lost compared to TS (0.536 vs 0.524), whereas a large amount of power is gained (0.564 vs 0.775) when there is an effect. A point worth noting: one might expect that such a loss in power for TS seen in Table 4.3 would be matched with a large gain in reward. But in this case since the effect size is relatively small (0.1), TS has a much higher proportion of optimal allocation (0.860) which does not translate strongly to reward gain (0.536), whereas TS PostDiff, by adapting the amount of uniform random allocation based on the size of the effect, has lower proportion of optimal allocation (0.738), thereby reducing reward slightly (0.524), but increasing power greatly relative to TS. We thus see the value of doing more UR assignment when effect sizes are small.

Recommendations for the Choice of Hyperparameter c

In order to give recommendations for the choice of the hyperparameter c, we first examine how the probability ϕ_t, given in equation (4.3), defined as the posterior probability of observing an arm-mean difference lower that a threshold c, evolves as more participants are seen. In Figure 4.6, we show this estimated probability for various values of c, for an effect size of both 0 and 0.1. We estimate ϕ_t as $\hat{\phi}_t = \hat{\mathbb{P}}(|\tilde{p}_1 - \tilde{p}_2| < c)$, with $\tilde{p}_1$ and $\tilde{p}_2$ draws from the TS posterior distribution of the arm means, and $\hat{\mathbb{P}}$ the empirical probability, i.e., the average number of times an absolute difference between these draws resulted to be less than c.

We can see that when c is above the true effect size, $\hat{\phi}$ is increasing with sample size towards 1. When c is less than the true effect size, $\hat{\phi}$ is decreasing towards 0. These are non-trivial results, as they indicate that the additional UR allocation of TS-PostDiff is able to overcome the bias induced by TS. For example, the above shows that the tendency of TS to lead to overestimating the size of the difference in arms does not prevent ϕ from converging to 1 when the effect size is 0.0.

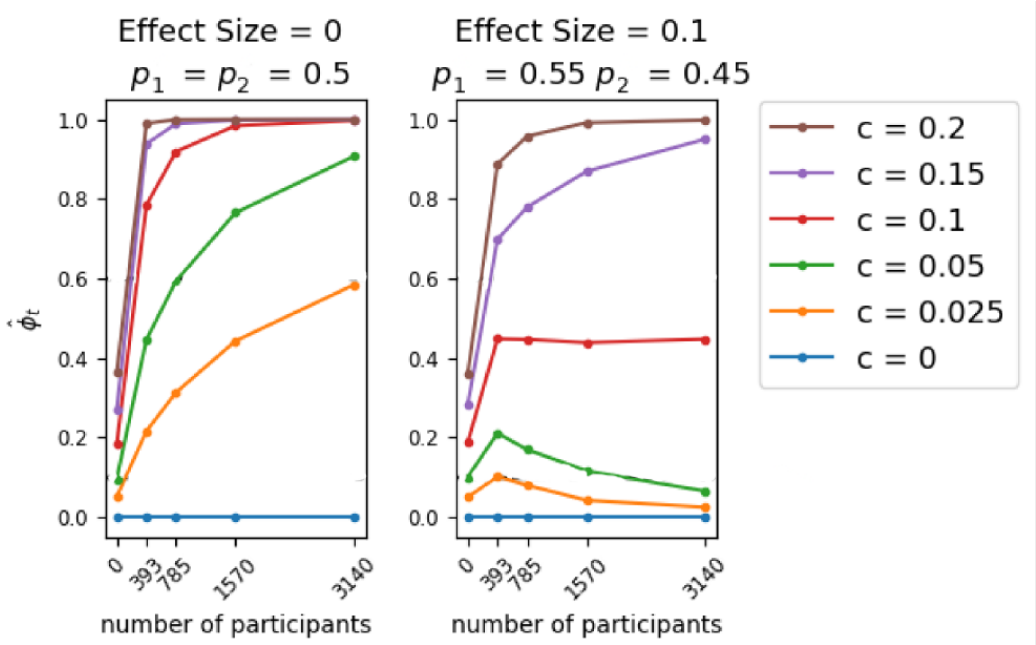

Figure 4.6. Estimated probability of observing an absolute difference between arm means less that a threshold c, i.e., $\hat{\phi}_t$, which also corresponds to the estimated probability of choosing actions uniformly, with respect to different thresholds c. Results are shown for different sample sizes and the two different effect sizes considered in this study: 0.1 (right plot) and 0 (left plot). We estimate ϕ_t as $\hat{\phi}_t = \hat{\mathbb{P}}(|\tilde{p}_1 - \tilde{p}_2| < c)$, with $\tilde{p}_1$ and $\tilde{p}_2$ draws from the TS posterior distribution of the arm means, and $\hat{\mathbb{P}}$ the empirical probability, i.e., the average number of times (out of a total of $10,000$ simulations) an absolute difference between these draws resulted to be less than c.

First Recommendation. Based on the this discussion of the behaviour of TS-PostDiff's ϕ_t probability trend with respect to c, the first recommendation would be to choose c as the value of an effect size which is small enough that we are willing to forgo reward in favor of improving data analysis capabilities. For example, if one is willing to accept an l loss in expected reward ($|p_1 - p_2|$) for a sub optimal allocation in favor of improved data analysis, Figure 4.6 tells as that if we choose $c = l$, when the true effect size is such that $|p_1 - p_2| < c$, we will converge to always using UR allocation, and if the true effect size is such that $|p_1 - p_2| > c$, we will converge to TS. In other words, if the true effect size is below what we are willing accept in expected loss in reward for a sub optimal allocation in favour of improved type-I error and power, TS-PostDiff will likely converge to choosing all arms with UR allocation. Similarly, when the true effect size is greater than what we are willing accept in expected loss in reward for a sub optimal allocation, TS-PostDiff will likely converge to TS.

Second Recommendation. If one isn't sure what reward they are willing to give up for improved data analysis, or perhaps such a decision would be easier to make if it was clearer what is being traded off, we advice to choose c equal to a guess for the true effect size. We motivate this recommendation based on results shown in Figure 4.7. Here, we illustrate how type-I error, power, and TS-PostDiff's percentage reward (compared to standard TS) for different values of

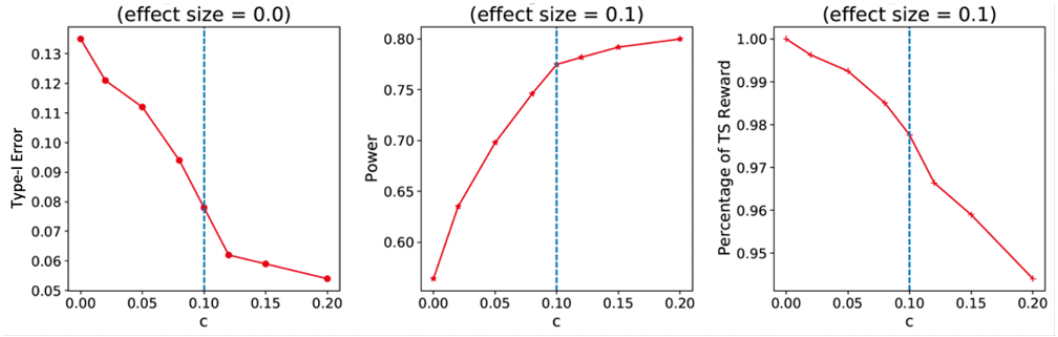

Figure 4.7. Type-I error (left; $p_1 = p_2 = 0.5$; $n = 785$), computed across 10,000 simulations. Different values of c are shown on the x-axis and indicators of performance are shown on the y-axis. Type-I error is computed as the percentage of simulations in which the null hypothesis is rejected when there is no difference between actions. Power is computed as the percentage of simulations in which the null hypothesis rejected when there is a difference between actions. Reward is computed as the percentage of total reward achieved by TS-PostDiff compared to the standard TS. These simulations show that as c increases, type-I error decreases and power increases, while reward decreases. Furthermore, improvements to power and type-I error diminish as c increases, while the to reward is roughly linear in c.

$c \in \{0.0, 0.025, 0.050, 0.075, 0.100, 0.125, 0.150, 0.200\}$, considering effect sizes 0 and 0.1 and a sample size of $n = 785$. The percentage of TS-PostDiff's reward is the average reward for TS-PostDiff for the given c value, divided by the average reward attained by TS; this metric allows us to see what percentage of the best possible (best of UR, TS, TS-PostDiff) reward we have achieved. This figure helps us discover a trend in the relationship between increasing c and these metrics. We see that we have diminishing returns in power when c exceeds the effect size of 0.1, whereas the percentage of TS's reward decreases roughly linearly. Type-I error also reaches diminishing returns, but a c value of 0.1 is close to the value where we see such diminishing returns. Choosing a c value of 0.1 is thus a reasonable choice, or in general a c value equal to the estimated effect size when effect sizes are smaller. If the effect size is larger though, then the loss in reward will be greater. Though the percentage of TS's reward is decreasing linearly, if TS reward is large, then this linear decrease will be more costly. Following the thresholds in behavioural science for small and medium effect sizes Cohen (1988), we don't recommend setting c higher than 0.2.

4.5 Discussion

Multi-armed bandit (MAB) algorithms have the potential to be extremely useful for conducting experiments in settings like healthcare, or other behavioural sciences, where a higher number of participants may benefit from the best interventions or arms. However, as the main goal of experiments is to draw conclusions about the arms, e.g., whether or not they differ from one another, this potential can only be realized if we are confident in our ability to draw correct conclusions. In this chapter, motivated by real-world educational experiments deployed with MAB algorithms - where instructors appreciated that the experiment had the ability to help their own

students, but were also concerned about whether the results reflected real differences in the quality of instructional materials (Williams *et al.*, 2016) - we evaluated the performances of hypothesis testing in such adaptive settings.

Through a simulation study, we showed that Thompson Sampling (TS), a common bandit algorithm, by searching for a superior arm with the aim to maximize reward, both tended to overestimate differences between arm means, with around 70% of simulations assigning the "superior" arm to more than 70% of participants (see Table 4.1), when the actual difference in mean was zero, and stopped sampling an arm even when there was still significant uncertainty about that arm's mean. These behaviors translated in a difficulty of drawing conclusions using statistical hypothesis testing: both with a frequentist framework (standard Wald Z-test and Welch's t-test) and a Bayesian framework (Bayes factor), we demonstrated that there was a high type-I error and a very low power when conducting inference on TS-collected data.

As a first exploratory approach, motivated by the typical data-assumptions in hypothesis testing, two alternatives for modifying the statistical test by incorporating knowledge about the dependent data collection process, were proposed. While the first alternative, using inverse probability weighting estimators to adjust the Wald Z-test, did reduce the bias in the estimated means, this was not enough to correct the inflated type-I error or decreased power (see Figure 4.5). With the second alternative, i.e., estimating the critical values used in hypothesis testing by simulating the TS-induced distribution of the Wald Z-test, the heightened type-I error was addressed, but at a significant cost in power (see Figure 4.5). As a whole, these results suggest significant barriers to using TS to conduct experiments, despite the fact that it has the potential to benefit participants in the experiment.

While the challenges of statistical hypothesis testing could be construed as only of interest to statisticians or behavioral researchers, they are important to consider from the machine learning perspective both for developing algorithms that can address scientists' real world challenges and for better understanding the behavior of typical bandit algorithms in cases where arms are equivalent. Many of the results we saw in the case study stem from two issues: a) the tendency of the algorithm to focus on a single arm even when both arms are equivalent, and b) failing to collect sufficient evidence to rule out the possibility that an arm that appears to be performing badly is in fact reliably worse than another. These tendencies cannot be counteracted solely by changing the way one analyzes the data, since the data look the same in some cases where there truly is a difference between arms and in cases where the arms are equivalent. They also cannot be counteracted solely by adding more participants and thus lengthening the horizon: when the two arms are equivalent, any pattern of sampling is equivalent in terms of regret, and the behavior of focusing on a single arm does not vanish asymptotically.

Motivated by the above, bandit algorithm modifications were explored. First, the novel problem of adding uniform random (UR) exploration to MAB algorithms, based on the estimated loss of reward in cases where there may exists or not a true arm-means difference, was introduced. Indeed, when the expected values of arms are similar, it is simultaneously true that the expected increase to reward from TS is small and sampling from both arms uniformly is important to increase the capacity to distinguish between arm means with a hypothesis test. Therefore, increases to power are more important and cost less reward when effect sizes are

small, so increasing UR exploration in those cases is particularly important. Based on this novel framework, a modified TS algorithm, TS-PostDiff, was developed. This was designed to increase UR exploration when there is either less evidence for a difference between arm means or evidence for a small difference. Overall, the proposed strategy resulted in a better balance of statistical power, type-I error and reward, compared to both UR and TS allocation (see Table 4.3).

Limitations. The current work was limited in exploring primarily the TS algorithm and considering a limited range of scenarios, including focusing on binary rewards and two arms. As the setting becomes more complex, we expect these challenges to persist, and further work is needed to explore the trade-offs between bandits being able to focus on better arms, perhaps permitting experimentation with a larger number of arms, and the inference concerns that we have laid out here. The case studies we developed can provide a foundation for what concerns to consider as well as highlighting the importance of considering cases in which arms are equivalent or nearly equivalent. In considering our hypothesis testing approaches, we focused on illustrating why small changes to the hypothesis testing approach do not address the inference issues satisfactorily. An illustration on how one can integrate machine learning knowledge into how statistical analysis is conducted, showing what worked, and what limitations persisted, was given. As the more straightforward modifications don't address the type-I error and power issues, this points to the need for further work on developing new statistical tests that consider the assumptions of particular algorithms for data collection (whether bandit algorithms or other methods like best-arm identification). While we show and believe that changes to algorithms like TS are needed to address these inference issues, there are still promising avenues for collaboration between the machine learning and statistics community to explore how changes in both the test procedures and the data collection might trade-off between best solutions from a regret standpoint and best solutions from an inference standpoint.

More generally, this work points to the need for formulating bandit problems that explicitly consider the quality of the evidence collected. Prior work has considered this in some bandit problem formulations, such as best-arm identification (Even-Dar *et al.*, 2002; Audibert & Bubeck, 2010; Russo *et al.*, 2018), power-constrained bandit algorithms (Yao *et al.*, 2020), and bandit algorithms that aim to correct for estimation error (Erraqabi *et al.*, 2017). None of these is explicitly concerned, however, with recognizing when arms are equivalent, an reward-maximization may become a secondary objective compared to "good-data quality" reliable inference.

Overall, there remains a great deal for future work to explore how to effectively collect data such that participants in experiments are able to benefit from accrued evidence, and the collected evidence is such that researchers can draw correct conclusions about the underlying properties of the arms. Here, we provided a foundation for future work on bandit algorithms that explicitly consider the reliability of statistical analysis in the objective. We hope that this may serve as motivation for the broader endeavour of bridging the gap between reward-maximising algorithms and scientific experiment design, by highlighting how balancing these competing objectives can be framed, and providing an instance of a solution.

Chapter 5

MHealth App to Promote Physical Activity in University Students: Results from A Micro-Randomized Trial[1]

Abstract

Low physical activity is an important risk factor for common physical and mental disorders. Physical activity interventions delivered via smartphones can help users maintain and increase physical activity, but outcomes have been mixed. Here we assess the effects of sending daily motivational and feedback text-messages in a multi-level micro-randomized clinical trial on changes in physical activity from one day to the next in a student population. We analyse 93 participants who used a physical activity app called *DIAMANTE* for a period of 6 weeks. Every day, their phone pedometer passively tracked participants' steps. They were micro-randomized to receive different types of motivational messages, based on a cognitive behavioral framework, and feedback on their steps. We use *generalized estimation equation* models to test the effectiveness of feedback and motivational messages on changes in steps from one day to the next. Sending any versus no text message results in an initial increase in daily steps (729 steps, $p = 0.012$) but this effect decreases over time. A multivariate analysis evaluating each text message category separately, shows that the initial positive effect is driven by the motivational, though the effect is small and trend-wise significant (717 steps; $p = 0.083$), but not the feedback messages (-297 steps, $p = 0.5$). Sending motivational physical activity text-messages based on a cognitive behavioral framework may have a positive effect on increasing steps, but this effect decreases with time. Further work is needed to examine the possibility of personalization and contextualization to improve the efficacy of text-messaging interventions on physical activity outcomes.

ClinicalTrials.gov Identifier: NCT04440553.

[1]Parts of the text of this chapter are extracted from the submitted/published manuscripts coauthored by the candidate and listed on page vii.

5.1 Introduction

Insufficient physical activity is one of the leading risk factors of death worldwide (WHO, 2018). It is associated with worse outcomes in many common chronic diseases, including diabetes, cancer, coronary heart disease and worse mental health outcomes, including depression and anxiety (Brugnara *et al.*, 2016). Additionally, a low level physical activity is associated with the occurrence of mental disorders such as depression (Choi *et al.*, 2019). The World Health Organization (WHO) recommends 2.5 hours of physical activity per week of moderate intensity (WHO, 2018). However, in 2018, almost half of American adults did not achieve this goal (CDC, 2020). Adolescents and university students show even lower levels of physical activity (Kriemler *et al.*, 2011; Castro *et al.*, 2020). Given the detrimental effects of low physical activity for individuals, families and society as a whole, there is a great need to develop effective interventions that help people to increase and maintain their physical activity patterns over time.

Behavioral interventions delivered via mobile devices, such as text messaging and/or smartphone apps, hold great promise for helping people engage in healthy behaviors such as increased physical activity. They may be able to overcome some of the barriers to physical activity, like lack of will power, and help to identify the benefits of, and opportunities for, exercise. They can also help users gain insight into their walking behavior and aid with goal setting and accountability. Literature reviews suggest that mobile interventions have beneficial effects on physical and mental health, with effect sizes up to 3.10 (Cohen's d; Cohen, 1988) after three month follow up (though null effects were also reported; Rose *et al.*, 2017; Roberts *et al.*, 2017; Rathbone & Prescott, 2017). Further, a meta-analysis showed that smartphone interventions led to an increase of 476.75 steps per day (Romeo *et al.*, 2019). In addition, a scoping review that included 30 studies also showed that physical activity interventions decreased depression and anxiety symptoms in young people (Pascoe *et al.*, 2020). Delivering interventions via mobile phones has advantages as it allows for a wide dissemination of these interventions, and provides a potentially low-burden and low-cost manner of support (Schueller *et al.*, 2019).

However, similar to *face-to-face* treatments, the effects of mHealth physical activity interventions are mixed (Stuckey *et al.*, 2017), and typically do not seem to be sustained over longer periods of time (Romeo *et al.*, 2019). In addition, a systematic review of physical activity interventions showed that 12 out of 20 reviewed interventions resulted in increased physical activity, but noted the level of evidence regarding the immediate and the long-term effects of interventions to promote physical activity among university students is limited. One reason is that these interventions are not personalized enough (Triantafyllidis *et al.*, 2019). For instance, despite collecting a wealth of data, they do not adapt their messaging strategy (content and frequency) to changing behavior of participants over time, or allow people to vary their goal. Further, because most mHealth studies evaluate the effects of the intervention as a whole, and not its separate components, much remains unknown about which components of smartphone interventions are effective to increase daily physical activity.

A state-of-the-art experimental design proposed for testing the proximal effects of the intervention components in mHealth is the micro-randomized trial (MRT)

design (Klasnja *et al.*, 2015). As introduced in Section 3.1, in an MRT, individuals are repeatedly randomized to different intervention options. This design allows researchers to test separate intervention components, such as different categories of text-messages (as in the current study), and explore short-term effects of intervention components (e.g., steps within a 24-hour period after sending a message).

In this study, we tested a mobile phone application that sends daily feedback and motivational text messages, consisting of different categories of messages (motivational ($k = 4$) and feedback ($k = 5$)) based on a cognitive behavioral framework: the *Capability, Opportunity, Motivation, Behavior* (COM-B) model (Michie *et al.*, 2011). COM-B is a behavioral change model that proposes that engaging in a particular behavior depends on the dimensions of capability (physical and psychological), opportunity (social and physical) and motivation (need to engage in the behavior more than in other behaviors). These are interacting dimensions, and interventions must target at least one of them to achieve behavior change. The model hypothesizes that changing perceived opportunities, capabilities and motivations could lead to long lasting behavior change. COM-B has previously been used to identify physical activity barriers (Flannery *et al.*, 2018), and to design physical activity interventions (Carney *et al.*, 2016; Nyenhuis *et al.*, 2017).

The primary aims of the study were: 1) to examine the overall effectiveness of a text-messaging mobile app for improving physical activity, defined as the change in users' steps counts from one day to the next one, and 2) to assess the effectiveness of different types of text-messages (motivational and feedback) on physical activity. Secondary aims included: 1) to understand if there is a time effect on the interventions effectiveness; 2) to explore whether the adaptive RL-based policy has an improved outcome in terms of physical activity compared to the static uniform random assignment of messages; 3) to examine participants' pre- and post-intervention depression, anxiety and behavioral activation scores and differences between random messaging and adaptive sampling.

5.2 Experimental Design

Mobile App. We employed the DIAMANTE mobile phone app (`https://diamante.healthysms.org/`), developed by *Audacious Software* and the authors (Aguilera *et al.*, 2020). This application tracks step counts by pooling from *Google Fit*, *Apple HealthKit* or the built-in pedometer on patients' phones. We use the *HealthySMS* text-messaging platform, developed by Audacious Software and the authors, to send text-messages and manage participant responses back to our system. The app only needs to be installed once, but has to remain open consistently. The app is designed in English and Spanish versions and is freely available as a download from the *Apple App Store* and *Android Google Play*.

Participants. Undergraduate and graduate students of University of California, Berkeley were recruited through the *Social and Experimental Research Lab* (Xlab) app. The Xlab app is advertised during campus events and online advertisements on Facebook. Students go through an online screening process to determine their eligibility. Students that did not have a smartphone, were not able to exercise due to disability, or had plans to leave the country during the study, or where

not between the ages of $18 - 65$, were not eligible to participate. The study was approved by the *Committee for Protection of Human Subjects* (ID: 2019-04-12118). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the *1964 Helsinki Declaration* and its later amendments or comparable ethical standards.

Study Visits. Before entering the study, participants came for a baseline study visit and informed consent. Participants filled in baseline survey measures of interest on Xlab computers using *Qualtrics*, including the measures outlined below, and participant demographics and information about current mobile technology familiarity and utilization. All participants received assistance if necessary in downloading a pedometer application onto their phone. Participants were instructed to have the app open at all times. They received 15 USD for their participation in the baseline visit, independently of whether or not they used the app and received text-messages. At six-week follow-up, participants were invited for an online remote exit interview on their personal digital devices, for which they received 25 USD.

Measures and Outcomes of Interest. At baseline and at six-week follow-up, students were asked to complete a survey which included questions about demographics, socioeconomic status, health status, physical activity and the following psychological questionnaires: the *Patient Health Questionnaire-8 items* (PHQ-8; Kroenke *et al.*, 2001), the *General Anxiety Disorder-7 items* (GAD-7; Spitzer *et al.*, 2006), the *Behavioral Activation for Depression Short-Form* (BAD-SF; Santos, 2013)) and the *International Physical Activity Questionnaire - Short Form* (IPAQ-SF; Lee *et al.*, 2011). PHQ-8 is a reliable and valid eight-item measure of depression severity over the past 2 weeks in clinical and general population samples (Kroenke *et al.*, 2001). The PHQ-8 omits the PHQ-9's suicidality question and was found to be preferable to the PHQ-9 in research settings and online studies (Spangenberg *et al.*, 2012). All the baseline information constituted the contextual or independent variables.

The main outcome of interest, representing physical activity, was chosen to be the number of steps. Steps were passively collected by the pedometer mobile phone application continuously during the time participants remained in the intervention, provided that they did not close the application. For each day, we calculated the total number of steps recorded between $00:00$ and $23:59$. We then computed the change in daily step count, defined as today's step count minus yesterday's step count, to study and model the improvement in physical activity. *Daily step change* represented the dependent or reward variable in this study.

Experimental Factors: Text-Messages. We adapted a text-messaging bank that we originally designed for the clinical population (Aguilera *et al.*, 2020) of the DIAMANTE Study (see Section 3.1), for the current population, by removing messages that talked about chronic disease or family. Messages were designed to fit into the three dimensions of the COM-B model. Further, about half the messages were framed with a social connotation (i.e. exercising with friends or being healthy for others), and half were individually framed (exercising for yourself). We additionally added messages about the benefits of walking on brain health and concentration. Besides the motivational messages, participants also received one feedback message

daily, at approximately the same time as the motivational messages (two minutes apart). The feedback messages contained information on step count and step goal in the previous day. See Tables 5.1 and 5.2 for examples.

Motivational category	**Example**
M0. No message	NA
M1. Capability/Self-belief	*Push yourself a bit further with the help of friends. They believe in you!*
M2. Opportunity	*Find* 30 *minutes in your day to go for a walk. That is less time than it takes to watch one episode of a TV show.*
M3. Motivation/Walk-benefit	*Going for a walk can improve your mood and clear your mind.*

Table 5.1. Motivational messages categories and examples within each category.

Feedback category	**Example**
F0. No message	NA
F1. Reaching goal	*Yesterday, you did not reach your goal.*
F2. Steps walked yesterday	*Yesterday, you walked* 3824 *steps.*
F3. Walked more/less than goal yesterday	*Yesterday, you walked more than your goal.*
F4. Steps walked yesterday, plus a positive/negative message	*You walked* 4000 *steps yesterday, you can do better!*

Table 5.2. Feedback messages categories and examples within each category.

Study Design. The design of this study is an MRT (Klasnja *et al.*, 2015). In each study day, interventions or treatments are defined by the full factorial design with a total of three factors representing *Motivational Messages* (M), *Feedback Messages* (F), and the *Time Frame* (T) when the message was sent, of $k = 4$, $k = 5$ and $k = 4$ levels each, respectively. One level of both M and F corresponded to a control treatment, i.e., no message sent. Each participant received a combination of M, F and T every day, constituting a multi-level MRT design (Xu *et al.*, 2020). These designs allow us to examine the effect of sending a message versus no message on physical activity and explore the effectiveness of different categories of messages.

In addition, for evaluating the adaptive RL-based strategy, we compared it to a static uniform randomization, in which the same types of messages were sent out randomly. Our initial design, which is currently applied to the main DIAMANTE Study (see Section 3.1; Figure 3.1), additionally planned to randomize participants to the uniform random vs the adaptive condition. Because of technical difficulties (errors in execution with incoming data), we only enrolled a subset of participants towards the end of the study (after the errors were fixed) in the adaptive group, starting from October 21st, 2019. Thus, the majority of participants were assigned to a uniform random combination group. However, participants were not aware of their group membership until after the study ended.

5.3 Adaptive RL-based Strategy

Every day, we implemented an adaptive learned decision mechanisms for deciding on: 1) the feedback message, 2) the motivation message and 3) the timing of the message. Each combination of levels of the three experimental factors represented an arm or action. To increase personalization, the decision about which message to send did also take into consideration the contextual variables: time-independent variables such as baseline socio-demographic information, and time-dependent covariates, including the day of the week (Monday-Sunday), the data steps of the previous day, and the number of days since messages from different categories were sent.

For this study, we adopted the randomized Thompson Sampling MAB strategy, previously described in Section 3.1.2. However, we proposed some modifications from its original version described in Algorithm 4, to account for the specific mHealth setting and the high dimensionality of the context.

Similarly to Algorithm 4, we assumed that the expected reward (i.e., the daily step change) is a linear function of the context-action feature $f(\mathbf{X}_t, A_t) \in \mathbb{R}^{d+1}$, i.e.,

$$\mathbb{E}[Y_t|\mathbf{X}_t, A_t] = f(\mathbf{X}_t, A_t)^T\boldsymbol{\beta},$$

with $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ an unknown reward parameter, Y_t the reward variable, $\mathbf{X}_t \doteq [\mathbf{X}_{0t}, \mathbf{X}_{1t}, \ldots, \mathbf{X}_{d't}] \in \mathbb{R}^{d'+1}$ the contextual vector of d' number of baseline covariates, and A_t the actions or experimental variables at time or day t, with $t = 1, \ldots, T$. The inclusion of covariate $\mathbf{X}_{0t} \doteq \mathbf{1}$ is considered in order to incorporate the model intercept. Note that the context-action feature can be any function of the contextual and action variables, that may have a relevance according to the behavioural scientist. It can include interactions between those variables, or other combinations that may account for specific characteristics of the data, such as *habituation*, which we will discuss later.

Regularization. Differently from Algorithm 4, which considers a Normal prior distribution, here, we used a Bayesian linear regression setting with a *Normal-Inverse-Gamma* (NIG) prior on the regression coefficients (mean and variance). This keeps coefficients small and minimizes overfitting and provides some regularization, by shrinking coefficients. Formally, for each day t, we assumed a Gaussian distribution for the reward, i.e.,

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

and a multivariate Normal-Inverse-Gamma conjugate prior for the joint distribution of the parameters vector $\boldsymbol{\beta}$ and the variance parameter $\sigma^2 > 0$, i.e.,

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

with $\boldsymbol{\mu}_\beta \in \mathbb{R}^{d+1}$, and a, $b \in \mathbb{R}_{>0}$ fixed and known prior hyper-parameters. Alternatively, assuming Σ_β known, the prior distribution can be formulated as

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

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

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

The resulting vector $\left\{\tilde{\boldsymbol{\beta}}_{(t)}, \tilde{\sigma}^2_{(t)}\right\}_{t=1}^T$, with $\tilde{\sigma}^2_{(t)} \sim \text{IG}(a^*, b^*)$ and $\tilde{\boldsymbol{\beta}}_{(t)} \sim \mathcal{N}_{d+1}(\boldsymbol{\mu}^*, \tilde{\sigma}^2_{(t)}\Sigma^*)$ provides samples from the joint NIG posterior distribution, while $\left\{\tilde{\boldsymbol{\beta}}_{(t)}\right\}_{t=1}^T$ and $\left\{\tilde{\sigma}^2_{(t)}\right\}_{t=1}^T$ provide samples from the marginal Normal and IG posterior distributions, respectively. Based on the posterior samples, at each iteration t the optimal arm $\tilde{a}_{(t)}$ will be the one that maximises the 'a-posteriori' estimated expected reward, $f(X_t, A_t)^T\tilde{\boldsymbol{\beta}}_{(t)}$, where the posterior nature is reflected in $\tilde{\boldsymbol{\beta}}_t$.

Habituation. In many real-world scenarios, temporal changes in the reward distribution structure are an intrinsic characteristic of the problem, and the stationary assumption, as in the linear TS without any additional time modelling considerations, may be simplistic. One may expect for instance that the effectiveness (reward) of a specific intervention on a user would deteriorate over time or continuous assignment to that intervention. It is, thus, unlikely that users' preferences will remain stable over time, and the collected data becomes progressively obsolete as the interest for the items evolve. In mHealth, or, more generally, in behavioural sciences, this phenomenon, known as *habituation*, is a recognized pattern and modelling issue.

Habituation is a form of behavioral plasticity and learning defined as a decrement in response as a result of repeated test stimulus (Bouton, 2007). In a text-messaging app as the DIAMANTE app, sending the same message category or combination (arm) repeatedly over time may eventually result in a decreased response of the app user, thus, a reduced reward over time. Consequently, an arm which was optimal for and individual over a certain number of initial days, might not be optimal anymore, and the modelling procedure should be able to detect this distributional change. This departure from the stationarity assumption, which has dominated much of the MAB literature so far, raises fundamental questions as to how one should model temporal uncertainty in rewards, and how to benchmark performance of candidate policies. Existing bandit proposals include *recovery bandits* (Pike-Burke & Grunewalder, 2019; Cella & Cesa-Bianchi, 2020), based on which the expected reward of each arm varies according to some (unknown) function of the time since the arm was last played; *rested bandits* (Gupta *et al.*, 2011), which assumes that the time varying reward probability follows a simple Brownian motion; *rotting bandits* (Levine *et al.*, 2017),

where each arm's expected reward decays as a function of the number of times it has been pulled.

We took a similar approach as in Levine *et al.* (2017) and modelled the outcome of interest (daily steps change) as a function of the number of times since a text-message category was not sent. Assuming again K different text-messages or arms, and a fixed number of days T, in each of which only one of the available categories can be sent. Now, for each text-message categories $A_j, j = 1, \dots K$ and day $t = 1, \dots, T$, we denote by $Z_{A_j,t}$ the number of days since category A_j was last played, where $Z_{A_j,t} \in \mathcal{Z} = \{0, \dots, Z_{\max}\}$ for a finite $Z_{\max} \in \mathbb{N}$. More specifically, at day $t+1$, we have that for each $A_j, j = 1, \dots K$,

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

$A_{j,t}$ is the dummy coded version of the message category representing whether category A_j was chosen at time t: $A_{j,t} = 1$ means that on day t category A_j was assigned ($A_t = A_j$), while $A_{j,t} = 0$ means that on day t a category different from A_j was selected ($A_t \neq A_j$). Note that, as $Z_{A_j,t}$ depends on the past selected categories, it is a random variable as well.

Let now $\overline{\mathbf{Z}}_t \doteq (\overline{Z}_{A_1,t}, \dots, \overline{Z}_{A_K,t}) \doteq (Z_{\max} - Z_{A_1,t}, \dots, Z_{\max} - Z_{A_K,t})$, for $t = 1, \dots, T$, be the vector of these derived auxiliary variables, which we call habituation or recovery context. The idea is that based on the closeness in time a certain text-message category was sent the reward might be positively or negatively affected. Particularly, with (negative) habituation, we refer to the case when sending the same text-message consecutively may cause habituation and loss of its potential effect in terms of step change. In this setting, a higher $\overline{Z}_{A_j,t}$ represents a higher degree of habituation at time t related to category A_j, indicating a higher loss in terms of reward if the same message category is going to be sent, while a zero value indicates that sending again that message is not going to be affected by a habituation phenomenon. We hypothesize, in line with the behavioural literature, that in a certain fixed number of episodes $Z_{\max}$, if a specific text-message is not sent, habituation due to that message will stop to occur.

Based on this reasoning, at time t, we model the daily step change of text-message A_j as a linear function of this arm, of the time and/or action invariant context $\mathbf{X}_t$ and of the related arm dependent contextual variable $\overline{Z}_{A_j,t}$, i.e.,

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

Thus, the expected reward of every arm changes at each round t, and this change depends on whether arm A_j was previously played and how many rounds ago. We also included in the model other time-dependent contextual variables, in addition to the time-independent baseline one, in order to account for the effect of time. Being a multi-factorial trial we also examined interactions between the different message categories and between them and the time variable.

5.4 Statistical Analysis

Descriptive statistics of interest of the sample were reported in terms of means and standard deviation for relevant continuous measurements and frequency and percentage for categorical variables.

Multivariable regression analysis. To test the effectiveness of interventional messages, we used the *Generalized Estimating Equations* (GEE) model (Liang & Zeger, 1986), a widely used longitudinal data analysis method in mHealth (Bolger & Laurenceau, 2013). GEE models represent an extension of generalized linear models and quasi-likelihood estimation methods (McCullagh, 2018), and are designed for analyzing clustered (or longitudinal) data which may be correlated within a cluster (in our case the user) but are independent between clusters. We used the `geepack` package in `R` (Halekoh *et al.*, 2006), and employed an independent working correlation structure (within clusters), taking into account the *Quasi-Information Criterion* (QIC) method (Pan, 2001). In addition, when there are time-dependent covariates, the GEE estimator was shown to be consistent under the independent working correlation structure, and thus it is recommended as a "safe" analysis choice (Pepe & Anderson, 1994).

Using GEE, we first examined the effect of sending any versus no message on step change. As a second step, we also looked at a model evaluating the effect of sending a feedback or motivational message, including the interaction between them, to understand if and to which extent one of these components has effect on step change. Third, we explored a model examining the effect of the different categories of feedback ($k = 4$) and motivational ($k = 5$) text-messaging categories. All models were adjusted for time (study day). In this study we primarily focused on the effect of the messaging system and examined the effect of the time of day factor only as a secondary aim.

In the main analyses, we included both the uniform random and Thompson Sampling group to have an increased sample size. However, in the Thompson Sampling group, after 2 weeks of micro-randomized messages, messages were no longer delivered randomly, but via a learned policy mechanism, which could alter results. Therefore, as a sensitivity analysis, we also re-ran all analyses while removing the Thompson Sampling group.

Using GEE, as a secondary objective, we also assessed differences in overall change in physical activity between the Thompson Sampling and the uniform random group, adding the study group as an independent variable. We also examined the effect of group membership on changes in PHQ-8 scores, GAD-7 scores and BAD-SF between baseline and follow-up using a two-way repeated measures ANOVA. However, these analyses were exploratory, as the groups were not randomized. Because we did not perform randomization, we also did not assess the influence of socio-demographic and baseline factors on the effect of the adaptive intervention (moderators of the intervention) as originally planned.

Missing data. Being the outcome of interest defined as the difference in the step count between two consecutive days, we first, excluded participants with less than 2 days of data. Then, a complete case analysis (based on the outcome variable of interest) was carried out for evaluating the effectiveness of text-messages. As a

sensitivity analysis, we also performed the analyses using missing-data imputation on the reward variable.

We used multiple imputation as missing-data imputation technique. More specifically, we employed multivariate imputation by chained equations, also called fully conditional specification or sequential regression multiple imputation (Azur *et al.*, 2011). This method has emerged in the statistical literature as one principle method of addressing missing data, and a dedicated package in R exists (Zhang, 2016). We used this package adapting the existing functions to our longitudinal data.

Power analysis. Being an exploratory study, power analysis did non represent a primary analysis of this work. However, in order to estimate an adequate sample size, we originally conducted our power analysis based on a repeated-measures analysis. For longitudinal data analysis one needs to conduct simulations and when we started this study we had no prior data available to do so. Based on previous literature, we expected that participants would show an increase of between 1000 and 2000 steps over the whole study period (Harries *et al.*, 2013; Walsh *et al.*, 2016; Fukuoka *et al.*, 2019). Thus, hypothesizing an increase from 6000 to 7500 steps with a standard deviation (SD) of 2000 steps, a correlation between measurements of 0.5, and aiming for a power of 90%, we would have need 76 subjects.

After the end of the study, and based on data from the current participants, we also applied the novel proposed GEE-type data power analysis method of Xu *et al.* (2020). This method, indeed, required the average standardized effect sizes of the intervention levels of motivational and feedback messages to calculate power. Even if conducting such a post-hoc power analysis is not advisable, using this novel method, specifically developed for mHealth MRTs, allowed as to obtain at least indicatively an estimate of the power we would have for the analyses looking at the message categories separately.

5.5 Study Results

Participants Data

The sample consisted of 103 students enrolled from September 12th 2019 to October 25th 2019, who did sign the informed consent, agreeing to participate to the study. Of these, 7 participants did not receive the text-messages due to technical issues (iOS updates, $n = 5$, or wrong language setting in the Google Play store, $n = 2$) and 3 received text messages, but never transmitted data back to our server. This left 93 participants for the analysis (see flowchart in Figure 5.1). Baseline characteristics of the sample included in the analysis ($n = 93$) are shown in Table 5.3.

Overall, $t = 670$ (16%) days of observation with missing steps were recorded and removed from the main analysis. On average, in 45 days of study, subjects received a motivational message (M1, M2 or M3) on 27 days and a feedback message (F1, F2, F3 or F4) on 30 days. System errors led to the messages not being sent out for 16% of the time, corresponding to the missing data amount. In these cases, subjects did not receive any messages. We have kept these non-randomized days ($t = 700$ days) in the main analysis, coded these as the no message category (M0 and/or F0).

This gives us a better understanding of the effect of the messages compared to not sending any message. However, also we conducted a sensitivity analysis (reported later in this section) removing the days in which subjects did not receive messages because they were not directly randomized.

Baseline Covariate; $n = 93$	
Gender; n (%)	
Female	65 (69.9%)
Male	27 (29%)
Other	1 (1.1%)
Age; Mean (SD)	20.2 (2.47)
Ethnicity; n (%)	
Asian or Pacific Islander	51 (54.8%)
Hispanic/Latino(a)	11 (11.8%)
Multi-ethnic	10 (10.8%)
White or Caucasian	19 (20.4%)
Refused	2 (2.2%)
Born in the US; n (%)	55 (59.1%)
Physical activity; n (%)	
Engaging in regular physical activity last 6 months	48 (51.6%)
Wants to be more physically active	88 (94.6%)
Minutes of moderate/vigorous exercise/week*; Mean (IQR)	150 [90, 171]
Psychological questionnaires; Mean (SD)	
PHQ-8 (depressive symptoms)	5.61 (3.62)
GAD-7 (general anxiety)	4.73 (4.84)
BADS-SF	31.1 (8.33)

Table 5.3. Baseline characteristics of $n = 93$ analysed participants data: Mean (SD; standard deviation) or Mean (IQR; interquartile range) for continuous variables, and n (%) for categorical variables. PHQ-8: Patient Health Questionnaire-8, GAD-7: General Anxiety Depression Scale-7, BAD-SF: Behavioral Activation for Depression Scale. *Measured by the International Physical Activity Questionnaire (IPAQ).

Multivariable Regression Analysis

Sending any text message versus no message. On days that any message was sent (e.g. individuals either received a feedback or a motivation message or both) versus no message, sending a message initially resulted in an increase in the change of number of steps by 729 ($p = 0.012$, standardized effect size $\delta = 0.147$). However, this effect diminished linearly over time (trend-wise significant), with a decrease of 33 steps on average for each additional study day ($p = 0.004$, $\delta = -0.007$, as shown in Table 5.4). Here, we only include time as a linear term, studying its main effect and interaction with the interventions (variables of interest). Additional analyses reported in Appendix G suggest the time's effect could be non-linear, meaning that at the beginning of the study there is an increase in the outcome Y, but after a while, increasing time will result in a decrease in Y (when the coefficient estimate

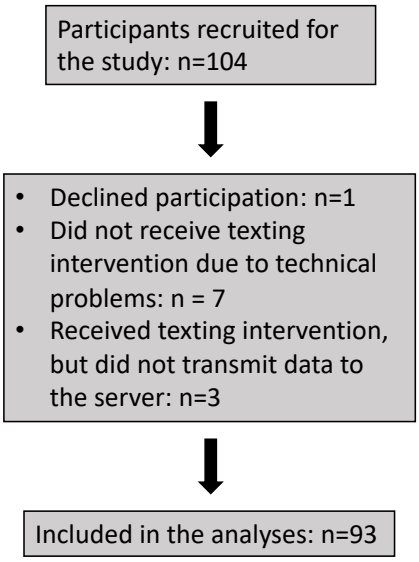

Figure 5.1. Flowchart of DIAMANTE participants enrollment with exclusion reasons.

is negative, as in our case). However, when including the non-liner term in the multivariable model with the intervention variables, results do not change.

Covariate	Estimate	95% CI	p-value
Message Sent	729	[163, 1295]	0.012
Study Day	27.4	[8.53, 46.4]	0.005
Message*Study Day	−33.2	[−56, −10.4]	0.004

Table 5.4. Results of the GEE model studying effects of sending any versus no message on steps change. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Differences between motivational and feedback message. Sending a motivational message trend-wise increased the change number of steps initially after correcting for time and interactions between message categories (717 steps; $p = 0.083$, $\delta = 0.144$, see Table 5.5).

Covariate	Estimate	95% CI	p-value
Motivation	717	[−93.6, 1527]	0.083
Feedback	−297	[−1089, 496]	0.463
Study Day	11.9	[−3.97, 27.9]	0.141
Motivation*Study Day	−14.6	[−44.7, 15.6]	0.344
Feedback*Study Day	−1.65	[−31.2, 27.9]	0.382
Motivation*Feedback	−24.4	[−676, 627]	0.570

Table 5.5. Results of the GEE model studying effects of motivational and feedback message on steps change. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Different categories of motivational and feedback messages. In an exploratory manner, we also examined the effect of the different types of feedback

and motivational messages. Overall, there was a trend of an increase in change of steps from one day to the next when participants received a motivational message, in particular a self-efficacy message (414 steps; $\delta = 0.083$, $p = 0.077$, see Table 5.6) that discusses the (mental) health benefits of physical activity, and an opportunity message (410 steps; $\delta = 0.083$, $p = 0.089$, see Table 5.6), that discusses finding opportunities for exercise. There was a significant decrease when subjects received a feedback message with the number of steps they walked yesterday (−665 steps; $\delta = -0.134$, $p = 0.002$, Table 5.6). After including all interactions, the positive effect of motivation remained significant, but not feedback.

Covariate	Estimate	95% CI	p-value
M1. Capability/Self-belief	253	$[-247, 753]$	0.320
M2. Opportunity	414	$[-45.4, 874]$	0.077
M3. Motivation/Walk-benefit	410	$[-63.2, 883]$	0.089
F1. Reaching goal	−147	$[-703, 408]$	0.603
F2. Steps walked yesterday	−406	$[-907, 95.5]$	0.113
F3. Walked more/less than goal yesterday	−126	$[-579, 327]$	0.585
F4. Steps walked yesterday, plus a positive/negative message	−665	$[-1082, -247]$	0.002
Study day	1.64	$[-2.58, 5.85]$	0.447

Table 5.6. Results of the GEE models studying the effects of different categories of feedback and motivation on steps change. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Uniform Random vs Thompson Sampling group. Twenty-seven participants (29% of the overall sample), enrolled after the October 20th 2019, received messages chosen by the adaptive Thompson Sampling algorithm. Group membership (uniform random or Thompson Sampling) did not have a significant effect on steps (−515, CI $= [-1536, 506]$, $p = 0.32$). The Thompson Sampling group was small compared to the uniform random group (causing an imbalance between the two groups) and we did not conduct a group randomization due to technical difficulties. RCTs with a longer follow-up time therefore must be conducted to assess the added benefit of Thompson Sampling for increasing physical activity. We are currentlyconducting a larger RCT to examine this in a patient population (Aguilera *et al.*, 2020).

Psychological Questionnaire Scores

The internal consistency of each of the questionnaires administered to students showed acceptable-high reliability scores, as measured by the *Chronbach*'s α coefficient (Cronbach, 1951) at baseline assessment. A coefficient of 0.77, 0.93 and 0.80 was obtained for the PHQ-8, GAD-7 and BAD-SF, respectively. A total of 82 out of 93 subjects also provided follow-up data. PHQ-8 scores significantly increased from baseline to follow-up (5.67 (SD = 3.72) to 8.35 (SD 3.40), $p \leq 0.001$). There were no significant changes in anxiety (4.66 (SD 4.79) to 5.84 (SD 4.72), $p = 0.11$) and behavioral activation scores (31.3 (SD = 8.57) to 29.2 (SD = 8.60), $p = 0.12$). Changes

in scores did not differ between the uniform random and Thompson Sampling group (all p's ≥ 0.12).

Sensitivity Analyses

Multivariable regression analyses on the uniform random group only. As a first sensitivity analysis, we repeated all the analyses focusing only on the uniform random group ($n = 66$), for which stronger statistical properties are demonstrated being a uniform and independent rule of assignment, thus not subject to bias due to adaptive nature that characterizes the Thompson Sampling. We show and discuss more in depth this issue in Chapter 4.

Similarly to the main results, sending a message initially resulted in a positive effect on steps, but decreased over time. However, the effects were no longer significant which could be due to decreased power. The positive effect on steps seemed to be driven by motivational messages but this also lost significance after adding interaction terms. The main categories showing significance were again a opportunity message (a borderline $p = 0.06$ with an estimated coefficient of 517) and a feedback message with the number of steps participants walked yesterday plus a negative/positive feedback message (-541 steps, $p = 0.03$). All the results are reported in Appendix H.

Missing data imputation. As we performed multiple imputation considering a number of three imputations, we report results for all the imputed datasets (see Appendix I. This will allow assessment of the robustness of the method, as well a more exhaustive sensitivity analysis. Consistent with the main original findings (no data imputation), sensitivity results with data imputation show the same directions of effects. With respect to all the three different analysis we performed we can see that:

1. In the assessment of single message categories (different feedback and motivational categories) generally the effect of the same categories resulted to be statistically significant. The positive effects of the motivational messages M3 and M2, and the negative effect of the feedback messages F4 were still significant. In addition, imputing the missing data improved this significance, showing lower p-values, and an additional significant message category, i.e., F2.

2. In the assessment of feedback and motivation messages without specifying each category, we still had consistency in terms of direction of the effect, however out of the three imputed datasets, only one resulted in a significant effect of motivational message. The other two were not significant, but note that the p-value in the original dataset was 0.08, thus only a borderline significance, which is consistent with the average result of the three multiple imputations.

3. In the assessment of sending any message versus no message we still have a significant result in one of the imputed datasets, consistent with our original results. Our intuition behind the loss of significance in the other two imputed datasets related to the distribution of the imputed datasets compared to the

original ones: both absolute number of steps and step change had a higher variability in the imputed datasets.

Post-hoc power analysis. Based on data from the current participants, the effect-sizes for the motivational messages were respectively 0.0734, 0.121 and 0.1079, which can be categorized into small (i.e., < 0.2) standardized effect size. In order to achieve a power of at least 80% to detect these effect sizes, a minimum sample size 117, or a minimum study period of 57 days with 93 participants, is required. In conclusion, the current sample of 93 participants for 45 days would likely be sufficient to detect larger standardized effect sizes, e.g., above 0.108 for each of the message levels, but not small ones. This illustrates that we did not have sufficient power to detect these effects of individual message categories.

GEE vs Mixed Models. In mobile-health longitudinal studies, the two major and most common methods for analyzing data (whose potential correlation due to repeated measurements should be taken into account during the statistical analyses), are represented by GEE, which is a marginal (population-average) model, and the linear (or generalized) mixed model (LMM) with fixed and random effects, which is a conditional (subject-specific) model. The two approaches have thus different targets for inferences (population vs subject level) and address subtly different questions about longitudinal change. A fundamental difference between GEE and LMM is the interpretation of model's coefficients. Random effect models coefficients have subject-specific interpretation in terms of change in the transformed mean response for any individual. However, marginal models ignore such changes within subjects: here, the regression coefficient describes how the average rates for any variable may change in the study population.

In our study, we used the first approach (i.e., GEE) for several reasons. First, because our primary analysis focused on the population-level effect of interventions, not only for getting general insights on interventions effects, but also to derive reasonable prior specification (of our parameters of interest) to be used in the main DIAMANTE clinical study. Second, because several authors suggest that in general the estimation-equation approach provides a more useful approximation of the truth compared to mixed models (see e.g., Hubbard *et al.*, 2010), that interpretation from mixed models is more complicated in some settings Fitzmaurice *et al.* (2012), and that a greater variability in the estimates (due to the heterogeneity among individuals) characterises the latter. Finally, our preliminary results with an LMM model suggested that for our outcome of interest (i.e., change in the steps change), the random effect of study participants did not contribute to explain the total variability (variance of the random effect ≈ 0), but estimates' CIs were generally wider (see Table 5.7 for a comparison with model in Table 5.6). We used the `lmer` function of `lme4` (Bates *et al.*, 2014) `R`' package. The extent of this subject variation can thus be fully or virtually-fully explained by just the residual variance term alone, so there is no enough additional subject-level variation to warrant adding a subject-level random effect.

However, this is not true anymore when the absolute number of steps is taken as dependent variable: in this case considering subject-levels is clearly important, as they explain almost 50% of the variability in all the different multivariable models considered in this study. Our interpretation is that the change in the number of

Covariate	Estimate	95% CI	p-value
M1. Capability/Self-belief	253	$[-229, 735]$	0.304
M2. Opportunity	414	$[-51.4, 880]$	0.082
M3. Motivation/Walk-benefit	410	$[-54.7, 874]$	0.084
F1. Reaching goal	-147	$[-664, 369]$	0.577
F2. Steps walked yesterday	-406	$[-917, 106]$	0.121
F3. Walked more/less than goal yesterday	-126	$[-645, 393]$	0.634
F4. Steps walked yesterday, plus a positive/negative message	-665	$[-1183, -146]$	0.012
Study day	1.64	$[-11.8, 15.1]$	0.812

Table 5.7. Results of the Linear Mixed Model studying the effects of different categories of feedback and motivation on steps change. The random effect of study participants was included in the model. CI: Confidence Interval.

steps (from one day to another) variable, compared to the absolute daily number of steps, eliminates the individual baseline walking tendency.

For LMM, Qian *et al.* (2020) showed that standard software can be used to obtain a valid estimate of the fixed effects if the time-varying covariates are independent of the random effects parameters conditional on past history. When the time-varying covariates are allowed to be endogenous (e.g., depending on the outcome process or previous treatment assignments), estimation of the LMM fixed effect coefficients may lead to bias because it no longer corresponds to the conditional interpretation of the parameters (Pepe & Anderson, 1994). More recently, Hu *et al.* (2021) also examine the conditions under which the LMMs work in the presence of time-varying endogenous covariates, proposing an propose a variation of the LMM method that jointly estimates the fixed effects and the random effects under a ridge-type penalty on the latter.

5.6 Discussion

This study examined the effectiveness of motivational text-messages on changes in daily steps in a student population. We found that receiving any text message versus no message was initially associated with an increase in steps, but this effect was weakened over time. Looking at the effect of the message categories separately, a positive effect on physical activity seemed to be driven by the motivational messages broadly, and not the text-messages that provided feedback on participants' steps.

Of note, the effects we observed in the current study were small (Cohen's $d < 0.2$). A possible explanation for our small effects and overall lack in effect of our messages over time is that the messages were not contextually (i.e., adapted to participants' daily contexts) and not personally (i.e., altered to fit with a person's personality profile) tailored enough. For instance, previous studies showed that tailored messages, that adapt to time-varying factors such as time of day, day of week and participants' work schedules, are preferred by participants, and more effective than generic messages (Lee *et al.*, 2015b). For instance, Klasnja *et al.* (2015), using an MRT design, found that contextually tailored walking suggestions increased

bouts of physical activity. This MRT showed that sending multiple physical activity text messages a day increased steps with an average of 496 daily (Klasnja *et al.*, 2015). Comparable to our study however, the effect of sending messages decreased over the course of the study. Previous meta-analyses also showed that physical activity interventions seem to be effective only on the short term (Romeo *et al.*, 2019). Although we had a low percentage of drop out (3%), participants may have paid increasingly less attention to the messages as the study continued, leading to messages losing their effects. Engaging users in digital interventions, especially when they are unsupported, is one of the greatest challenges facing digital health today. One difficulty with relying on text messages is that it is not possible to confirm if individuals opened and read messages. Our previous work found that, when required to respond to text-messages, 25% of the sample was disengaged, e.g., did not respond to any of the messages (Figueroa *et al.*, 2020). Future work should specifically focus on how to measure and sustain engagement with texting interventions, that don't require a response, over time.

Our findings suggest that motivational messages may be more beneficial in increasing physical activity than feedback messages, though they lost their effectiveness over time and effects were small. Specifically, we found positive trend effects of self-efficacy, which revolves around the belief that one is capable of behavior change, and opportunity, the belief that one is capable of behavior change. Our results suggest that these two behavior change categories may be more important than messages about the benefits of exercise. University students, an educated sample, may already be aware of the beneficial effects of exercise, thereby these messages may be less impactful. Messages based on the COM-B framework may show some effectiveness in motivating individuals for physical activity and could be a suitable component within a physical activity intervention. Of note, when we removed days that participants were not randomized, the positive effect of motivational messages on step change was no longer significant. This may be due to reduced power to detect small effects. Our results need to be confirmed by future work.

We found that feedback on individuals' steps (whether they reached their goal or not the previous day) may have no effect on step change, or even has a negative effect. Results from systematic reviews showed that feedback should be actionable (e.g. when, how and where can you exercise to reach your goal today) to be effective (Schembre *et al.*, 2018). Future interventions may benefit from providing concrete actions a participant can undertake to reach their daily personal goals, in addition to feedback on their number of steps or goal achievement. For example, an app may provide information on how many steps individuals still need to walk that day to reach their goal, and give them personalized suggestions, such as places to exercise.

We did not observe any differences between the participants who received micro-randomized messages, and participants who received messaging chosen by a reinforcement learning policy. The duration of our study may have been too short for the algorithm to effectively start learning, especially given the limited number of participants. Further, Thompson Sampling presents with challenges including modeling the right reward function (outcome), low learning speed when data is sparse and assessing the usefulness of contextual variables (Tewari & Murphy, 2017; Liao *et al.*, 2020). To date, mHealth studies using ML have shown some promising

effects, but the number of studies is still too small for a rigorous evaluation of the machine learning methods (Triantafyllidis & Tsanas, 2019).

It is important to note that our current multi-level micro-randomized trial design (MLMRT) is intended to evaluate the effects of intervention components (in this case types of messages), but not the effectiveness of the intervention as a whole. The Thompson Sampling group was small ($n = 27$) and we did not conduct a group randomization due to technical difficulties with the Thompson Sampling algorithm. RCTs with a longer follow-up time therefore must be conducted to assess the added benefit of Thompson Sampling for increasing physical activity. We are currently conducting a larger RCT to examine this in a patient population (Aguilera *et al.*, 2020).

The nature of the college semester may also help to explain our finding of increased depression scores from baseline to follow up. This finding was unexpected, since previous work has found beneficial effects of physical activity interventions on mental health in students (Pascoe *et al.*, 2020). At the end of the study participants might have experienced greater stress due to upcoming exams and thereby increased depression symptoms. This may be another factor that led to reduced effectiveness of the messages over time. Importantly, our findings underline that students are a population particularly vulnerable for mental health issues, as corroborated by previous work41. Students may need more tailored mental health support within physical activity interventions to cope with the pressures of college life, especially during exam periods.

Limitations. This sample consisted of a convenience sample (university students), with various levels of baseline physical activity. Although the majority of the participants indicated a wish to become more physically active, they might not all have primarily joined the study to increase their physical activity. Thus, they may be an under-motivated sample. Further, this study suffered from technical issues such as system errors leading to not sending out messages and missing participant steps because Internet connectivity was needed to pull participants' steps from their phones. Additionally, participants might not have always carried their phones, leading to additional days with an unreliable estimate of steps. Further, our power analyses showed that our sample size was too small to detect significance of the individual levels of motivational and feedback messages (which was an exploratory analysis). In addition, even though young adults tend to carry their phone with them for most of their waking hours (Atas & Çelik, 2019), using phone built-in pedometers are sensitive to errors, particularly when participants forget their phones (Duncan *et al.*, 2018), or carry their phones in their pocket (Silva *et al.*, 2020). Further, we examine the effects of the messages on steps over a 24 hour period. Because messaging times were randomized, we expect that this is a well founded approximation of the effect of the message on the daily steps. However, ideally we would have examined the effect of messages within a certain number of hours. We collect data using in-built phone pedometers to increase real world validity, but because of this we did not have the fine grained data available, like the number of steps per minute. The nature of our design therefore does not allow us to assess with certainty if the change in steps was propelled by the message, or if there are carry-over effects of messages received in the previous days.

Chapter 6

RL for Optimizing MHealth Applications: Lessons Learned and Guidelines for Design Decisions[1]

Abstract

Providing behavioral health interventions via smartphones allows these interventions to be adapted to the changing behavior, preferences and needs of individuals. This can be achieved through Reinforcement Learning (RL), a sub-area of machine learning. However, many challenges could affect the effectiveness of these algorithms in the real world. We share our experience and provide some guidelines for decision-making. Using thematic analysis, we describe challenges, considerations and solutions for algorithm design decisions, in a collaboration between health services researchers, clinicians, statisticians and data scientists. We use the design process of an RL algorithm for a mobile health study, *DIAMANTE*, for increasing physical activity in underserved patients with diabetes and depression. Over the 1.5-year project, we kept track of the research process using collaborative cloud Google Documents, Whatsapp messenger and video teleconferencing. We discussed, categorized, and coded critical challenges. We group challenges to create thematic topic process domains. Nine challenges emerge, which we divide into 3 major themes: 1) Choosing the model for decision-making, including appropriate contextual and reward variables; 2) How to deal with missing or incorrect data in real-time; 3) Weighing the algorithm performance vs effectiveness/implementation in real world settings. The creation of effective behavioral health interventions does not depend only on final algorithm performance. Many decisions in the real world are necessary to formulate the design of problem parameters to which an algorithm is applied. These considerations and decisions must be documented and evaluated before and during the intervention period, to increase transparency, accountability and reproducibility.

[1]Parts of the text of this chapter are extracted from the submitted/published manuscripts coauthored by the candidate and listed on page vii.

6.1 Introduction

Mobile health applications (apps) such as smartphone and text-messaging interventions have proven effective in eliciting beneficial health outcomes, including better mental health management (Firth *et al.*, 2017a,b), weight loss (McCarroll *et al.*, 2017) and increased physical activity (Gal *et al.*, 2018; Murray *et al.*, 2017; Roberts *et al.*, 2017). However, only a small percentage of users use behavior change apps over a long period of time (Baumel *et al.*, 2019), and mobile applications have not become part of routine medical care (Lyles *et al.*, 2019; Rowland *et al.*, 2020). Engagement is particularly low for unsupported interventions (Weisel *et al.*, 2019).

One explanation for low retention and declining effectiveness of apps, is that they are not responsive enough to users' changing needs. More recently, an increasing interest in machine learning to optimize digital behavioral health interventions has emerged. Delivery via smartphones allows to personalize these interventions to individuals' preferences and needs (Nahum-Shani *et al.*, 2018). Using machine learning, the content of interventions can also dynamically adapt with changes in participant behavior over time, to maximize outcomes (Triantafyllidis & Tsanas, 2019).

Reinforcement Learning (RL), a subfield of machine learning, is a powerful method to use in various healthcare settings because it can optimize sequences of decisions (Rabbi *et al.*, 2019; Yu *et al.*, 2019b). Several studies have started using RL for optimizing the delivery of text-messaging (Piette *et al.*, 2015; Liao *et al.*, 2020). For example, using simulations, Piette *et al.* (2015) showed that RL to optimize text-messaging for medication adherence could produce a $5-14\%$ increase in adherence, could predict medication barriers, and detect when messages were sent too frequently. RL algorithms for mobile health make predictions based on incoming participant data, and use these to make decisions for individuals (e.g. what message should the participant receive and when). As more information is collected over time, the algorithm improves its predictions, hence makes more effective decisions. A previous mobile health study elucidated that RL algorithms can learn new strategies over time to maximize physical activity (Yom-Tov *et al.*, 2017). The algorithm altered its decision making strategy when participants changed their exercise behavior (e.g. walked less) because the weather worsened (Yom-Tov *et al.*, 2017).

However, the use of RL presents multiple challenges in the real world. A systematic review on machine learning in mobile health (beyond solely RL) identified only a few Randomized Controlled Trials (RCTs), the highest level of evidence in clinical medicine (Sibbald & Roland, 1998), and a lack of studies in clinical practice settings (Triantafyllidis & Tsanas, 2019). Guidelines for designing and using these algorithms in clinical settings are needed. For instance, as opposed to simulation studies, clinical studies may have unforeseen difficulties such as data errors, involve a large interdisciplinary team, and need to be executed within a limited timeframe.

Outlining the challenges and decisions to make throughout the process of algorithm development for a clinical RCT increases transparency, replicability and likely outcomes of behavior health studies using RL. Further, identifying and solving issues related to differences in scientific strategies of disciplines will help to increase the productivity of interdisciplinary collaborations (Lach, 2014).

We recently started the *Diabetes and Mental Health Adaptive Notification Track-*

ing and Evaluation (DIAMANTE) RCT: a smartphone application that uses RL to optimize physical activity text-messaging in underserved patients with diabetes and depression (Aguilera *et al.*, 2020).

We analyzed study notes of a 1.5-year design process of implementing an RL algorithm: a collaboration between computer scientists, behavioral scientists, physicians, psychologists and statisticians. We discuss challenges of developing these algorithms for mobile health in real world settings and the solutions we implemented. This provides guidelines for decision-making, which can be used by other clinical and healthcare researchers.

6.2 Material and Methods

The DIAMANTE Study. DIAMANTE is a 6-month physical activity text-messaging study, which sends individuals motivational text-messages to help them increase their physical activity. The DIAMANTE study is a randomized controlled trial with three groups (uniform random, reinforcement learning, and a control group). The study is registered on clinicaltrials.gov: NCT03490253 and a brief introduction was given in Section 3.1 (see Figure 3.1). Phone-pedometers passively collect daily step counts on participants' personal phones. The study recruits low-income English and Spanish speaking patients with depression and diabetes served in a safety net setting. In a user centered design period, we developed our motivational messages based on a *Cognitive Behavioral Framework: Cognition, Opportunity, Motivation and Behavior* (COM-B; Michie *et al.*, 2011). This process included qualitative interviews, usability phases and crowdsourcing to categorize the messages. Participants indicated liking the final set of messages in pilot phases. However, they disliked messages that were perceived as repetitive, and suggested that personalized advice would make the content stronger. In these phases we identified issues that may harm user engagement related to missing data due to internet connectivity problems, server errors in sending out messages, and participants' low technical skills to access the app on their phone and transmit their data.

Experimental factors and RL algorithm. Participants receive two messages daily at approximately the same time. In the reinforcement learning group, the algorithm evaluates each morning which messages, and delivered within what time period, will likely increase steps for every participant in the upcoming day. In the UR group, participants receive the same messages, but they are micro-randomized: message categories and timing are delivered with equal probabilities instead of chosen by a learning algorithm. Micro-randomization is different from regular randomization where participants are randomized into intervention groups. In micro-randomization, interventions (here text messages) are repeatedly (here daily) randomized within participants.

The algorithm chooses the types of messages from different categories, their frequency and delivery time period. The action space is defined by a $5 \times 4 \times 4 \times 2$ factorial design: 5 intervention options for a *feedback* message and 4 intervention options for a *motivational* message, including the "no-message" category, 4 different time frames, and 2 social categories (individual or family). We assumed that our reward variable, i.e., the daily change in steps, is a linear function of contextual

variables, action variables and interactions between actions and action-contextual variables. The model contains contextual variables for each participant. These include time-fixed variables, such as demographics and clinical characteristics, and time-varying variables, such as day of the week. There were 71 contextual variables for each participant. Categorical contextual variables have been represented with dummy variables (binary variables) as well, while all non-binary contextual variables (such as age) have been normalized to a value between 0 and 1. Algoritms details have already been reported in Section 5.3.

Data sources to identify challenges and solutions. To describe challenges and solutions, we used study documents. Over the two years of the project, the team kept track of the research process using a combination of *Google Docs* tools, *Whatsapp* messenger communication and video conferencing (*Zoom/Skype*). Further, the team convened to discuss, categorize, and code critical challenges and solutions over the course of the study on numerous occasions. Field notes were taken, and audiotapes of key discussions were recorded and subsequently transcribed and coded. Challenges were grouped to create thematic topic process domains. Here, we review the challenges encountered and the solutions devised when creating the DIAMANTE algorithm.

6.3 Results: Potential Challenges and Solutions

Over the 1.5 year time period, the behavioral science team (Behavioral scientists, physicians and psychologist from the *University of California Berkeley and San Francisco*) met once per week. Additional meetings happened weekly with the App developer from January 2019 onwards. Starting February 2019, the data science team (computer scientists and statisticians from the *University of Toronto*, the *National University of Singapore* and the *Sapienza University of Rome*) met with members of the behavioral science team on a weekly basis until April 2020.

We coded 119 pages of notes and 82 pages of exported WhatsApp conversations. Further, we transcribed around 7 hours of data of key meetings, using an online automatic transcription service (*otter.ai*) combined with our own transcription where the automatic transcription failed.

Nine challenges emerged which were divided into 3 major themes. We discuss them below.

1. Choosing a learning algorithm. Standard RL algorithms may perform poorly with the limited data collected in mobile health studies: treatment (here text-messages) is provided up to a few times per day. We chose algorithms for contextual multi-armed bandit (MAB) problems: a problem of deciding which arm of an experiment to try, when the goal is maximizing reward from a distribution with unknown parameters. A detailed description and analysis of these algorithms is reported in Section 3.1.2. We used these algorithms based on our previous work and because they might be particularly effective for mobile health (Rabbi *et al.*, 2019; Tewari & Murphy, 2017). MAB algorithms simultaneously attempt to acquire new knowledge by exploring the different intervention options (here text-messages), and optimize decision based on acquired knowledge (e.g. which text-message led to a

positive reward before). Each intervention option is associated with a different reward function, also depending on participant's context: here participant characteristics (e.g. age, gender) or daily time-varying variables (e.g. day of the week).

We proposed as adaptive strategy the contextual linear Thompson Sampling algorithm illustrated in Section 3.1.2, Algorithm 4; and we decided to implement it after first two weeks in which text-messages were sent uniform randomly (analogous to an initial "burn-in" period, or, more appropriately, an "internal pilot" for acquiring some prior data). The TS choice was motivated by several reasons. First, its empirical and theoretical properties have been well-studied, showing great performances (Chapelle & Li, 2011; Agrawal & Goyal, 2013). Second, it is computationally efficient, thus particularly suitable for online learning (Russo *et al.*, 2018). Third, it represents a stationary randomized algorithm, and as such, its expected cumulative regret never grows linearly with time, which may happen in any deterministic stationary strategy (Russo *et al.*, 2018). Finally, TS has been widely applied in real-world applications, including mHealth (Liao *et al.*, 2020), showing successful results also with small amounts of data (Agrawal & Goyal, 2013).

MABs have been applied in fields including education research (Williams *et al.*, 2017) and mobile behavioral interventions including physical activity (Rabbi *et al.*, 2015) and sleep (Daskalova *et al.*, 2020). Using simulations, we previously showed that contextual MABs for educational technology interventions resulted in better student outcomes than non-contextual MABs or randomization (Shaikh *et al.*, 2019).

MAB algorithms have limitations. For instance, they are slow to adapt to changing circumstances due to external factors, including the weather or new illness (Rabbi *et al.*, 2019). Additionally, because they take only short-term rewards into account, they are not optimal for maximizing long-term outcomes. However, we expect that slight increases in walking from day to day, may result in changes in habits (Rabbi *et al.*, 2019), and thereby increases in overall steps over the duration of the study (6 months).

2. Variable selection. Contextual MABs formalize the reward model as a function of both intervention and contextual variables, which can be used for personalization. Thus, an adequate choice of the variables to be included in the model is crucial for unbiasedly estimating parameters of interest and causal relationships. As shown in Fig. 3.1, the DIAMANTE study collects a wealth of contextual variables at baseline, but guidelines for choosing variables to include in the model are lacking. The action space results in a high-dimensional space, where each arm is a combinations of all the available factor levels. This is complicated by the presence of a high number of both baseline and time-varying covariates, which may also interact with the interventions.

In the absence of reliable estimates at the start of the study (not enough data from pilot phases), to avoid missing potential important variables, we initially considered all arms, all baseline variables shown to be relevant in the literature, and included also action-action and action-contextual interactions. Given this high-dimensionality, we adopted a slightly different reward model from the one proposed in Algorithm 4, which may provide regularization by shrinking coefficients, and avoid overfitting (Marquardt & Snee, 1975). The full model is illustrated in Section 5.3.

In addition, during the study, we do evaluate the model every 3 months to improve performance through an iterative process. This includes assessing if we should remove

certain terms from the regression model, i.e. based on high correlations, or choose a different type of regression method (i.e. LASSO-related regression (Park & Casella, 2008), which automatically select variables by removing predictors from the model). We expect that algorithm tweaks become less necessary as the study progresses.

3. Choosing the reward variable. There are no guidelines for choosing the reward variable: the model's outcome/feedback. Some studies using RL algorithms use a reward of a 30-minute step count to increase short bouts of physical activity. However, in the current study, we only send out messages at one time point per day. In testing phases of our study, users indicated that receiving messages in the morning motivated them to be active later in the day, varying with their schedules. Therefore, we chose as proxy outcome variable for measuring physical activity the number of steps (collected by the pedometer on participants' personal phones). However, in order to account for users' baseline walking propensity, we decided to consider the steps change from one day to another, within a time interval of 24 hours after an intervention was sent, corresponding to the minimum amount of time between two consecutive intervention options. Compared to the steps count, the steps change measure also shows a closer Gaussian shape, assumed by our reward model. This is a trade-off, as this longer time period could introduce noise, but a short time window may miss meaningful activity. The reward thus depends on the purpose of the study, as well as the wearable instrument. Future work should evaluate algorithm performance for various types of reward variables.

Dealing with missing data in the reward variable. Many reinforcement-learning algorithms that are deployed in the real world were trained on very big datasets. In contrast, the average mobile health study has no more than 200 people. Further, there is a high risk of missing or unreliable data because participants may forget to carry their phones or their phones are not transmitting data. There are no best practices on how to deal with these missing values (e.g. omit them from the data-set, impute values etc.). There is a lack of mHealth studies which addressed this problem, and, in an online experimental setting it is particularly relevant as it may impact subsequent selection of interventions when reward is missing.

In our user testing phases, we noticed that technical errors led to participants at times receiving messages that faultily stated that they had walked 0 steps. Our app is unable to pull steps if the app is not consistently open in the background. Further, if someone is not connected to the Internet (for instance, if they are out of data or have their phone in airplane mode), the server is unable to pull the steps. As such, the algorithm would also treat this measurement as a 0-step count, which could lead to faulty decision-making of the algorithm. In addition, receiving faulty steps was frustrating for participants. We therefore decided on the following approach to be carried out online (during the interventions delivery):

- Don't send users any feedback messages if we detect 0 steps;
- Code the 0 steps as NA in the training data set. The NA coding also avoids potential bias which may arise in the estimation process when the step count (or reward) is null due to missing data and not actual zero steps;

- Use the *last observation carried forward* (LOCF; Hamer & Simpson, 2009) technique for the primary analysis for dealing with missing reward data in this case;
- Send participants a message ("*Diamante has not received your activity, so we'll be unable to send any messages. Please open your app*") asking them to go into the app;
- Contact participants by phone if we have not received any steps for more than 2 days.

As a sensitivity analysis, we also include multiple imputation of missing data, as described in Section 5.4. However, we take an exploratory approach and perform this analysis offline (after the collection of the data), rather that executing them online and allowing the bandit algorithm to use the imputed data. Future work should evaluate missing data imputation in the online learning process.

5. Addressing algorithm errors in real time In user-testing phases, we ran into multiple technical errors. For instance, the app did not always collect steps and messages sent by our online server were occasionally inconsistent with messages registered in our data export file. To prevent these errors throughout the study we took the following approach:

- The researchers receive an automated daily log of data errors (such as 0-step measurements;
- Throughout the study, we compare the data export (which messages participants received according to our analysis dataset) to both the messages logged on our online server and the messages we receive ourselves, through our enrollment in the program as a continuous internal test;
- We conduct preliminary checks of the data quality every 3 months. These consist of checking for missing values, assessing micro-randomization errors (e.g., check the uniform assignment in the UR group) and evaluating consistency among collected data.

6. Speeding up learning with limited time available. The rate of algorithm learning slows with sparse data. In our study, we only enroll approximately $10 - 15$ participants per month – and even fewer during the COVID-19 pandemic. To speed up the learning of the algorithm, it is recommended to have prior knowledge to inform the prior distributions of the algorithm. As this study was not designed with a pre-period of large scale data collection, we started the adaptive assignment only after an initial uniform random assignment of two weeks, for every participant in the RL arm. This approach was demonstrated to effectively speed up the TS learning, with priors informed by the acquired data, through simulation studies. Indeed, uniform random data can be used to create informed prior distributions for the TS condition. This way, algorithm decisions are based both on prior knowledge and incoming data, which increases the speed of algorithm personalization (Russo *et al.*, 2018). We lacked recommendations on how long to collect uniform data before the learning phase. Based on our experience, most participants who drop out do so in

the first month. Therefore, we settled on only two weeks of uniform randomization before switching to TS, to maximize the probability that most participants will receive RL for at least 2 weeks. In simulation studies using pilot data from our user design phases ($n = 10$), we confirmed that TS has a lower bias in estimation of reward with an initial UR policy.

7. Choosing a comparison group. We originally planned to compare RL to an arm with fixed content derived from the *National Diabetes Prevention Program* (Albright & Gregg, 2013). In this case, participants receive one message a day with a predefined content for a period of 6 months. Although this approach was most comparable to existing health education interventions, we decided to cleanly evaluate the effect of RL by changing the comparison group to uniform random. Both arms receive the same types of messages, but uniform random picks the messages with equal probability. The RL condition dynamically adapts treatment, allowing us to assess if sending messaging using a RL decision making algorithm is superior to choosing message categories and timings at random. In addition, we had a third control group who downloaded the app on their phone, but did not receive any messages except for a weekly mood message. This way, we could also compare the effect of our messages to not receiving any messages.

8. Evaluating algorithm performance within an implementation study. The DIAMANTE study is a hybrid effectiveness and implementation trial, set up to ensure the effectiveness of the app in a real world setting. Hybrid designs combine effectiveness and implementation research to reduce the time from initial concept to a working product (Curran *et al.*, 2012). This is important because traditional designs often result in efficacious interventions that are not effective in real world studies, particularly for digital interventions (Mohr *et al.*, 2017).

Our team navigated between making decisions to optimize algorithm performance, and maximize usability. For instance, sending message more than once a day with shorter reward periods may improve algorithm performance (Liao *et al.*, 2020; Yom-Tov *et al.*, 2017), but may also decrease user engagement (Eysenbach, 2006). Because our target users did not frequently use apps or texting, and were therefore at higher risk for dropout (Figueroa *et al.*, 2020; Avila-Garcia *et al.*, 2019; Nouri *et al.*, 2019), prioritized decisions that would benefit user engagement, where possible.

To further examine clinical effectiveness and implementation, we implemented a three-arm trial design, including RL, uniform random and a control group to balance the need of evaluating algorithm performance vs effectiveness and implementation of the overall intervention. Doing so, we will also be able to evaluate the effect of our text-messages irrespective of the use of RL, by comparing the uniform random (micro-randomized) messages to the control group.

9. Limited time for algorithm development within the context of a clinical trial. Typical RL-algorithm research involves several preparation phases, which improve algorithm performance but may take years to complete before a clinical trial. Here, we were only able to employ a preparatory phase of 9 months before the start of the RCT. Most of the limited years of study funding were dedicated to the clinical trial. Given the limited amount of time for research on algorithm performance, we: 1) conducted simulation studies for debugging and defining model parameters; 2)

ran preliminary quality data checks on preliminary collected data from a different population study (see Chapter 5); 3) analyzed the algorithm data from the study team members and an initial small group of participants in a preliminary study. This testing revealed crucial errors, which we could fix before deploying the algorithm.

6.4 Discussion

We described the challenges, decisions and solutions of designing RL algorithms to personalize mobile health applications in real world settings, using the design process of the DIAMANTE physical activity study as a motivating example. Qualitative analyses from 1.5 years of study notes showed that the most important decisions and challenges were related to the choice of the model and design variables for decision making, handling missing data and algorithm errors in real time, and maintaining a balance between intervention implementation and optimal model performance. These issues need to be taken into account, and documented, during the design process of RL algorithms for clinical studies.

Approach with multiple phases. Despite a limited time to develop our RL algorithm, we employed a multi-phased approach. This included simulations studies, pilot user testing and testing of the mobile application within our own study team. We recommend that all researchers using RL or other types of machine-learning to personalize digital health interventions employ multiple phases, which can happen simultaneously, in preparation of a rollout with clinical participants. This is crucial for identifying and fixing algorithm errors. Because of the complexity of the design process, developing these models requires a multidisciplinary team, with both deep technical expertise and profound knowledge of the clinical population of interest. Further, the RL algorithm design process should not stop when the clinical study starts. Instead, decisions should continuously be evaluated and algorithm development should work through an iterative process. Frequent data checks, automated reports about missing data, and internal testing are essential throughout the study in its entirety.

Micro-randomization. Another important lesson is that an RL study must have an adequate comparison condition, allowing to disentangle the performance of the algorithm, and the characteristics of the digital intervention overall on clinical outcomes. Such a comparison, as we choose here, is contrasting RL to uniform random, in which messages are micro-randomized. This will allow us to quantify the benefit of "adaptive tailoring" using RL. We also used a period of uniform randomization for all participants to inform priors for the algorithm, with the aim to speed up learning with sparse data, which is an important consideration for mobile health. Researchers may also choose to include a period of micro-randomization in order to determine decision rules (Klasnja *et al.*, 2015), e.g. at when to send messages based on participants' availability (not performed in this study).

Increasing engagement through RL. The type and delivery frequency of text-messages will adapt over time throughout the study based on data collected from participants every day. This learning algorithm aims to maximize the outcome, increases in steps, learning and updating over time based on incoming data. More

commonly, digital health studies only tailor their content in a user centered design process to the needs, wishes and norms of a group of individuals. In this study we both tailor the content as a whole through a UCD process, and the text-messaging delivery on a daily basis using RL. We hypothesize that this approaches increases participant engagement and thereby will be more effective. We will assess participant engagement by examining the times they accessed the app and read the messages, how they rated the app's usability, and their qualitative opinions.

The importance of developing guidelines for machine learning in mobile health A framework paper discussing the machine learning literature in health argued that the field lacks transparency, clear reporting, exploration for ethical concerns, and demonstrations of effectiveness (Vollmer *et al.*, 2020). While several studies have discussed RL algorithm performance for mobile health, for example in simulations, less discussed all the steps needed to develop these algorithms for clinical studies. Because this is a novel field, machine learning algorithms used in applied health settings often undergo less scrutiny compared to other clinical interventions (Vollmer *et al.*, 2020). Additionally, because of the excitement around artificial intelligence, some have warned that digital medicine must avoid a crisis of reproducibility like found in other biomedical fields (Stupple *et al.*, 2019). Recent RCT reporting guidelines for AI studies for clinical decision-making have begun to emerge (Liu *et al.*, 2020). Here, we provide guidelines specifically for the algorithm design part of mobile behavioral health studies.

We recommend that all studies using machine learning to optimize digital health interventions document their decision making process and identify critical issues and challenges they encountered. This further avoids the "black box" problem of not knowing how and why algorithms are making decisions.

Issues around choosing a model for decision making also need to be explored more. Here we chose an algorithm for contextual multi-armed bandit problems, as this algorithm may be particularly suitable for mobile health studies. There is a lack of research that compares the effectiveness of different RL models and assesses what kind of problems within mobile health they should be applied to.

Similarly, we choose changes in daily steps as our reward. Other physical activity studies using an RL algorithm have also used 30-minute steps after participants received a motivational message (up to 5 times per day to increase short bouts of physical activity; Liao *et al.* (2020)), and the increase in activity since the last motivational text-message (Yom-Tov *et al.*, 2017). Algorithm performance with various reward functions also needs to be explored. Further, here we measured steps using participants' pedometers on their personal phones to facilitate real world implementation. Using wristband accelerometer like fitbits may however to more reliable data. Future implementation work should explore whether the use of wristbands to measure physical activity is sustainable in low-income populations.

Strengths. To our knowledge, this is the first study to describe RL algorithm design decisions in the context of a multi-disciplinary collaboration for mobile health clinical studies in the real world. Notably, we conduct this work in a low-income ethnic minority population for whom the greatest health disparities exist, yet where novel methods are not often designed (Figueroa *et al.*, 2020; Allen & Christie, 2016; Schueller *et al.*, 2019). This work brings the challenge of balancing decisions that

boost algorithm performance, and those that maximize usability. Clinical studies with low-income populations bring challenges related to data errors, low-tech skills of users, working within a large interdisciplinary team, and a limited timeframe. Many of these issues are less relevant in simulation studies or work with convenience samples. This paper can be used as a framework of considerations for other interdisciplinary teams working, or wanting to work, in this space.

Limitations. Researchers working with other populations and/or health problems may encounter issues not described here. Further, we provide a framework of decision-making, but because this is an evolving field, we cannot be certain that our choices are optimal. Detailed analyses on the final dataset, in combination with results from other studies, will provide these answers. Finally, here we did not discuss consideration such as privacy and algorithm bias in detail, but these issues must be further explored. Finally, user ratings and experiences of the content are not included in the current algorithm. Future work should focus on incorporating engagement into RL algorithms. Because user engagement can be quantified in many different ways, future guidelines should also define consistent engagement measurements for studies to be comparable.

Conclusion. Creating effective behavioral health interventions using RL involves many decisions beyond evaluating algorithm performance. These considerations need to be documented and evaluated before and during the intervention period to increase transparency, accountability and replicability. As the application of machine learning into digital healthcare interventions increases, we need effective collaborations between different disciplines to do this work well in real-world settings.

Chapter 7

General Discussion and Conclusion

Reinforcement learning represents a powerful solution in a variety of healthcare domains where problems have a sequential nature, and optimal decision making requires a continuous interaction with the underlying domain's process.

The first relevant domain deals with developing evidence-based adaptive interventions (AIs). Notably, as illustrated in Section 3.1, an extended body of statistical literature has proposed RL as an alternative framework to standard statistical techniques in AIs, starting from the pioneering works of Murphy (2003), Robins (2004) and Murphy (2005b). While these originated within causal inference for developing and estimating optimal dynamic treatment regimes (DTRs), more recently, an increasingly active interdisciplinary area of research, consisting of computer scientists, behavioural scientists and statisticians, started to use RL for constructing optimal just-in-time adaptive interventions (JITAIs), i.e., AIs characterized by a continuous learning with interventions tailored to users' in-the-moment context or needs.

The second healthcare domain we identified, relates to the use of RL and multi-armed bandit (MAB) strategies to adaptively design clinical trials - this is discussed in Section 3.2. One example is given by response-adaptive randomization (RAR) designs, which also have a long history in statistics starting with Thompson (1933)'s work, but are gaining in popularity only recently due to their recongnized potential for improvements in cost and efficiency over traditional designs (FDA, 2019). They use data of previous cohorts to adapt allocation of patients in succeeding cohorts; if a treatment showed more promising or informative results in prior patients, the probability of being assigned to that treatment increases (Hu & Rosenberger, 2006).

Given the multidisciplinary nature of the use of RL in healthcare, the first aim of this thesis was to provide a comprehensive review of the state-of-the-art RL methodologies in the different applied domains, among which we identified the area of developing AIs and designing adaptive clinical trials (CTs) as the most effervescent. We started by formalizing the theoretical foundations of RL in a mathematical/statistical way, rather than the typical computer science characterization. Then, in contrast with the current existing surveys, which focus exclusively on either DTRs or mHealth JITAIs, or single areas of CTs designs, such as RAR, here, we provided a unified view of all these areas, incorporating DTRs and JITAIs under the

unique area of AIs. We reported their main divergences and analogies, reviewing both methodological and applied aspects.

From a methodological aspect, we noticed that in DTRs, the use of RL has been extensively evaluated, and several surveys exist (Chakraborty & Moodie, 2013; Chakraborty & Murphy, 2014; Tsiatis *et al.*, 2019). However, their clinical application is still very limited; the majority of the existing studies use real-world data as motivating or illustrative examples only. On the other hand, we identified an opposite tendency in the development of JITAIs in mHealth. Here, the majority of the surveys refer to real-world applications, generally targeting a specific problem area, rather than a methodological work or review, which currently does not exist, and in this work we aimed to fill the gap. In this case, we recognize that the application area, mostly related to behavioural aspects rather than clinical, might have less concerns in terms of treatment costs (compared to CTs), and the general goal might be more focused on optimizing a proximal (behavioural) outcome, made possible by the technological sophistication which allows managing and delivering interventions in real time, without additional costs per intervention.

To illustrate the potential of JITAIs and RL in the emerging area of mHealth, we reported details and results of a behavioural micro-randomized trial (MRT) we carried out for promoting physical activity in a preliminary population of university students (see Chapter 5). Using a dedicated mobile app, i.e., the DIAMANTE app (Avila-Garcia *et al.*, 2019), we examined the effectiveness of motivational text-messages on changes in daily steps, and we found that receiving any type of text message was associated with an increase in the number of steps. This study allowed us to investigate the challenges and the decisions a practitioner has to face when developing mHealth AIs. We discussed them in Chapter 6, along with the solutions we proposed for both the design of the MRT and the design of the adaptive RL-based algorithm. Qualitative analyses from 1.5 years of study notes, showed that the most important decisions and challenges were: choosing of the model and design variables, handling missing data, and maintaining a balance between intervention delivery and optimal model performance. All these issues suggest that a higher synergy between applied and theoretical sciences may improve the current state-of-the-art.

When the goal of an experiment goes beyond the pure reward maximization, typical of many mHealth studies and the most common RL problem, other challenges may arise, and these are particularly relevant in the context of medicine and CTs. Taking up on the previously mentioned RAR designs, several RL classes of algorithms such as MAB algorithms, have been demonstrated to be extremely useful for adaptively assigning patients to interventions, in a way that an increasing number of patients is randomized to the treatment that shows the best evidence-based outcome. However, given that the main goal of randomized CTs is to draw conclusions about the compared treatments, e.g., whether or not they differ from one another, this potential benefit of, can only be accepted if we have guarantees in terms of statistical inference and results generalizability. At the moment, there's a lack of both insights into how this algorithms perform in adaptively collected data, as well as which statistical and inferential solutions should be adopted in this settings.

Motivated by the lack of insights and reliable guarantees for hypothesis testing in adaptively collected data with MAB methods, through a simulation study in a 2-treatments setting (see Chapter 4), we showed that drawing conclusions using

standard statistical hypothesis testing is a major problem: both with a frequentist framework (standard Wald Z-test and Welch's t-test) and a Bayesian framework (Bayes factor), the resulted type-I error was remarkably inflated and power highly reduced, compared to an uniform random assignment. These results are consistent with findings from other studies (Villar *et al.*, 2015a; Bowden & Trippa, 2017; Zhang *et al.*, 2020c; Yao *et al.*, 2020). In an attempt to provide deeper insights into this issue, and possibly propose solutions, we investigated the behaviour of MAB strategies in single simulation trajectories and showed that common MAB algorithms such as Thompson Sampling (TS), by searching for the best treatment with the aim to maximize reward, tended to overestimate differences between treatments, stopping to assign an inferior treatment, even when there was still significant uncertainty about its mean outcome. This was true in cases of small difference between treatments, as well as in cases when there were no underlying difference at all.

Based in the findings of our simulation studies, we explored two alternatives for modifying the statistical test, by incorporating knowledge about the dependent data collection process. However, while an improvement in the type-I error control was reached, a terrific cost in terms of power was paid, meaning that these RL-based adaptive data collection tendencies cannot be counteracted solely by changing the way one analyzes the data. Motivated by this ineffective test statistic adjustments results, bandit algorithm modifications were explored. First, we introduced the novel problem of adding uniform random (UR) exploration to MAB algorithms, based on the estimated loss of reward in cases where there may exists or not a true treatment-means difference. Indeed, when the expected values of treatments are similar, it is simultaneously true that the expected increase to reward from TS is small and sampling treatments uniformly increases the power of detecting a difference between treatments' outcomes. Therefore, increases to power are favoured, at the cost of less reward when effect sizes are small. Then, based on this novel framework, we proposed a new version of TS, designed to increase UR exploration when there is less evidence for a difference between treatments. The proposed strategy resulted in a better balance of statistical power, type-I error and reward, compared to both UR and standard TS allocation. Overall, these results and analysis suggest that there remains a great deal for future work to explore how to effectively collect data such that participants in experiments are able to benefit from accrued evidence, and the collected evidence is such that researchers can draw correct conclusions about the underlying properties of the compared interventions.

In conclusion, we hope that this work, by highlighting how balancing competing objectives can be framed, and providing instances of solutions, may serve as motivation for the broader endeavour of bridging the gap between reward-maximising algorithms, such as RL methods, and scientific and generalizable knowledge. We also strongly believe that RL offers a powerful solution in these areas, and we hope that our contributions may incentivize a higher synergy and cooperation between statistical and machine learning communities for supporting applied clinical or behavioural domains in carrying out real-world studies that may improve the quality of interventions delivery. We also recognize that this cooperation is very timely due to the spread of mHealth applications and adaptive experimentations, which need to come with *trustworthy* and reproducible results in order to advance scientific progress and knowledge.

Appendices

A Marginal Structural Models with IPW

MSMs, originally proposed for estimating the effect of static treatment regimes (Robins, 2000), provide a powerful alternative to SNMMs for describing the causal effect of a treatment (hence "structural"), and pertain to population-average effects ("marginal" over covariates, baseline and time-varying, and/or intermediate outcomes). Differently from the conditional approach of SNMMs, which models the causal effect of a final blip as a function of the entire time-varying history (conditioning on that), the marginal approach of MSMs assumes models for the expectation of a potential outcome under a specified unobserved DTR $\mathbf{d}$, marginalizing over the covariate history $V_{\mathbf{d}} = \mathbb{E}_{\mathbf{d}}[Y] = \mathbb{E}[Y^{\mathbf{d}}]$, or alternatively as a function of the baseline covariates X_0 only, i.e., $V_{\mathbf{d}}(X_0) = \mathbb{E}_{\mathbf{d}}[Y|X_0] = \mathbb{E}[Y^{\mathbf{d}}|X_0]$. Most often, $V_{\mathbf{d}}$ is specified as a linear combination of components of $\mathbf{d}$, e.g., $\mathbb{E}[Y^{\mathbf{d}}] = f(\mathbf{d};\boldsymbol{\theta}) = \alpha + \boldsymbol{\theta}\mathbf{d}$, with $\mathbf{d} = (d_0, \ldots, d_T) = (a_0, \ldots, a_T)$ the full treatment history, $Y^{\mathbf{d}}$ the potential outcome that the subject would have observed under $\mathbf{d}$, and $\boldsymbol{\theta}$ a set of parameters. However, recently, more flexible, spline-based models have been considered (Xiao *et al.*, 2014).

Of the different available methods, including maximum likelihood (Daniel *et al.*, 2013) or targeted maximum likelihood estimation (Rosenblum & Van Der Laan, 2010) that have been proposed to estimate MSMs, or their parameters $\boldsymbol{\theta}$, IPTW (Robins, 2000; Neugebauer *et al.*, 2012) is the most commonly used. IPTW estimation attempts to control for confounding through assigning each participant a weight. The basic form of this weight for subject i at time t takes the form

$$w^{\pi}_{t,i} = \frac{1}{\prod_{\tau=0}^{t} \pi_\tau(A_{\tau,i}|\mathbf{H}_{\tau,i})},$$

where $\pi_t(a|\mathbf{h}) = \pi_t(A_{\tau,i} = a|\mathbf{H}_{\tau,i} = \mathbf{h})$, so that the denominator is the probability that the subject received the particular treatment history they were observed to receive up to time t, given prior observed treatment and covariate histories.

Applying the terminal weights $w_{T,i}$ to each subject in the sample results in a pseudo-population in which treatment is no longer affected by past covariates, breaking the confounding; but crucially, the causal effect remains unchanged. Then the parameters of the MSM coincide with those of the re-weighted observational marginal model, which may be estimated using standard methods on the re-weighted data. The resulting estimates are consistent under correct specification of the MSM and non-zero denominators. Overall, MSMs estimation is typically performed in two stages: in the first stage treatment weights are calculated; in the second stage the outcome model is fitted.

B Examples of Real-World DTRs Studies using RL

Reference	Domain (Actions)	RL Method	Data Source
Zhao *et al.* (2009)	Chemotherapy dosage	Q-learning with SVR & ERT	ODE model
Hassani *et al.* (2010)	Chemotherapy dosage	Q-learning	ODE model
Ahn & Park (2011)	Chemotherapy drug-scheduling	AC	ODE model
Humphrey (2017)	Chemotherapy dosage	Q-learning with CART, RF & MARS	ODE model
Padmanabhan *et al.* (2017)	Chemotherapy dosage	Q-learning	ODE model
Zhao *et al.* (2011)	Cancer therapy, time	Q-learning with SVR	ODE model based on real data
Fürnkranz *et al.* (2012) Cheng *et al.* (2011)	Chemotherapy dosage	PI	ODE model
Akrour *et al.* (2012) Busa-Fekete *et al.* (2014)	Chemotherapy dosage	IRL & PS	ODE model
Vincent (2014)	Radiation therapy scheduling	Q-learning, TD(λ), SARSA(λ), PS	Linear model ODE model
Tseng *et al.* (2017)	Radiation dose escalation	Q-learning with DL	Retrospective data
Jalalimanesh *et al.* (2017b)	Radiation dose & fractionation	Q-learning	Agent-based model
Jalalimanesh *et al.* (2017a)	Radiation dose	Q-learning	Agent-based model
Goldberg & Kosorok (2012)	Cancer treatments	Q-learning	Linear model
Yauney & Shah (2018)	Chemotherapy Radiotherapy dosage	Q-learning	ODE model

Table B.7. RL-based studies for developing optimal DTRs in Cancer. ODE: Ordinary Differential Equation; SVR: Support Vector Machine; DL: Deep Learnig; AC: Actor Critic; ERT: Extremely Randomized Trees; MARS: Multivariate Adaptive Regression Spline; CART: Classification And Regression Tree; RF: Random Forest; IRL: Inverse Reinforcement Learning; PS: Preference Learning; TD: Temporal Difference; SARSA: State-Action-Reward-State-Action; PI: Policy Iteration

C Existing RL-based R Packages for Developing DTRs

R Package Name	Functions & Methods
`DynTxRegime` (Holloway *et al.*, 2020)	`owl` (Outcome Weighted Learning; Zhao *et al.* (2012)); `bowl` (Backwards Outcome Weighted Learning; Zhao *et al.* (2015); `rwl` (Residual Weighted Learning; Zhou *et al.* (2017)); `qLearn` (Q-Learning Algorithm; Murphy (2005b)); `iqLearn` (Interactive Q-Learning; Laber *et al.* (2014b)); `optimalSeq` (Augumented Inverse Probability Weighting; Zhang *et al.* (2012b, 2013))
`DTRreg` (Wallace *et al.*, 2020)	`method = "gest"` (G-estimation; Robins (2004)); `method = "dwol"` (Dynamic Weighted Ordinary Least Squares; Wallace & Moodie (2015); `method = "qlearn"` (Q-learning; Murphy (2005b))
`iqLearn` (Linn *et al.*, 2015)	Interactive Q-Learning (Laber *et al.*, 2014b)
`qLearn` (Xin *et al.*, 2012)	Q-Learning (Murphy, 2005b)
	GGQ (Ertefaie & Strawderman, 2018) - code in Supplementary material
	V-learning (Luckett *et al.*, 2020) - code based on `optim` function
	Bayesian Machine Learning (Murray *et al.*, 2018) - code in Supplementary material

Table C.8

D Bayes Factor Computation in a Two-Arms Binary-Reward Setting

As discussed in Section 4.3.1, in the context of Bayesian inference, hypothesis testing can be framed as a special case of model comparison where each model refers to a hypothesis (Rubin, 1978). We assume we have two competing hypotheses, each of which corresponds to a separate model: H_0 is the model for the null hypothesis, which here is that there is no difference in arm means, and H_1 is the model for the alternative hypothesis, which here is that there is some difference in arm means. Bayesian hypothesis testing specifies separate prior probabilities $\mathbb{P}(H_0)$ and $\mathbb{P}(H_1)$ for each hypothesis. Assuming that we do not have any prior knowledge on which hypothesis may be more plausible, we take $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 0.5$. Then, based on the observed data $\mathcal{D}$, we quantify the evidence in favour of the model H_0 and compare it to the evidence in favour of model H_1 (or alternatively the evidence against H_0). For a given hypothesis H, under the Bayesian framework, this evidence is given by the combination of the likelihood function for the observed data, say $\mathbb{P}(\mathcal{D}|\theta, H)$, which depends on an unknown parameter θ, with each of the prior distributions of the unknown parameter.

We give below details on the likelihood, priors and parameter of interest in our specific two-arms binary setting. For each of the hypothesis-specific models, averaging (i.e., integrating) the likelihood with respect to the prior distribution across the entire parameter space yields the probability of the data under the model and, therefore, the corresponding hypothesis. This quantity is more commonly referred to as the marginal likelihood and represents the average fit of the model to the data. The ratio of the marginal likelihoods for both hypothesis-specific models is known as the Bayes factor, whose general formulation is given by

$$BF_{10} = \frac{\mathbb{P}(\mathcal{D}|H_1)}{\mathbb{P}(\mathcal{D}|H_0)} = \frac{\int_{\Theta_1} \mathbb{P}(\mathcal{D}|\theta, H_1) \times \mathbb{P}(\theta_1|H_1) d\theta_1}{\int_{\Theta_2} \mathbb{P}(\mathcal{D}|\theta_2, H_0) \times \mathbb{P}(\theta|H_0) d\theta_2},$$

with $\mathcal{D}$ being the observed data, and Θ_1 and Θ_2 the parameter spaces for model 1 and 2, respectively.

In the context of a two-arm case with binary rewards, the Bayes factor may be computed in closed form by assuming a Binomial model for the reward and a Beta prior for the unknown parameters of the Binomial distribution. More formally, denoting again with A the arm and Y the reward, we model the conditional reward for each arm as $Y|A = k \sim \text{Binom}(1, p_k)$ and assume Beta prior distributions for the unknown parameters p_ks, with $k = \{1, 2\}$, i.e., $p_k \sim \text{Beta}(\alpha_k, \beta_k)$. Conjugacy allows us to compute the posterior distribution of the parameter in closed form. After having observed a sample of size $n = n_1 + n_2$, with n_1 and n_2 the sample size of arm 1 and arm 2, respectively, and $S_1 = \sum_{i=1}^{n_1} Y_i$ and $S_2 = \sum_{i=1}^{n_2} Y_i$ the number of successes in each group, we have that $p_k|\mathcal{D} \sim \text{Beta}(\alpha_k + S_k, \beta_k + n_k - S_k)$, for $k = \{1, 2\}$. Here, we consider a non-informative Beta prior distribution for both arms, with $\alpha_1 = \alpha_2 = 1$ and $\beta_1 = \beta_2 = 1$, equivalent to a Uniform distribution in $[0, 1]$. Comparing the competing hypothesis of no arms difference (H_0) and an actual arms difference (H_1) requires comparing the marginal likelihood of a model for which the parameters p_1 and p_2 are the same (thus, the arm means have a

common probability p, i.e., the success rate distribution does not depend on the arm group) and the marginal likelihood of a model for which the parameters p_1 and p_2 are different (thus, each arm group has a different success rate). This leads to a pooled prior and then pooled posterior distribution in case of model H_0, and to the product of two separate priors and then two separate posterior distribution for model H_1. More specifically, we have that

$$\mathbb{P}(\mathcal{D}|H_0) = B(\alpha_1 + \alpha_2 + S_1 + S_2, \beta_1 + \beta_2 + n - S_1 - S_2)$$
$$\mathbb{P}(\mathcal{D}|H_1) = B(\alpha_1 + S_1, \beta_1 + n - S_1) \times B(\alpha_2 + S_2, \beta_2 + n - S_2),$$

where B denotes the Beta function.

Comparing these two quantities will give us the Bayes factor BF_{10}, whose value will give us the evidence against the null hypothesis H_0, similar to hypothesis testing in the frequentist setting. We consider two cutoffs for "rejecting" the null-hypothesis: i) a threshold of 1, which is the more intuitive cut-point for selecting one of the two models, ii) a threshold of 3. The choice of 3 is based on Jeffreys' scales of evidence for model selection (Jeffreys, 1961; Kass & Raftery, 1995), which considers a $BF_{10} > 3$ as substantial evidence in favour of the alternative hypothesis.

E Sensitivity to Priors for TS

Figure E.1 presents results comparing Jeffreys' prior to Beta$(1, 1)$ prior in the prior choice for the Beta-Binomial model of the TS algorithms. Jeffreys' prior is a non-informative prior (Jeffreys, 1961) which we use to assess the sensitivity of the algorithm with respect to type-I error and power. In the Bernoulli reward setting Jeffreys' prior is equivalent to a Beta$(1/2, 1/2)$. As we can see in Figure E.1, Jeffreys's prior doesn't influence results substantively: the difference between Beta$(1, 1)$ and the Beta$(1/2, 1/2)$ Jeffreys' prior with respect to type-I error is no greater than 3% (regardless of hypothesis test). Also, the difference in power doesn't differ by more than 5% (regardless of hypothesis test). Further prior sensitivity analysis for TS is conducted by Rafferty *et al.* (2019). In particular, they examine the settings of prior for an arm having a mean which is below, between, and above that of the true arm expected reward in the context of two other statistical test. Priors above encourage more exploration due to being more optimistic, whereas priors below encourage less. The result is that having an optimistic prior will improve power and reduce type-I error somewhat.

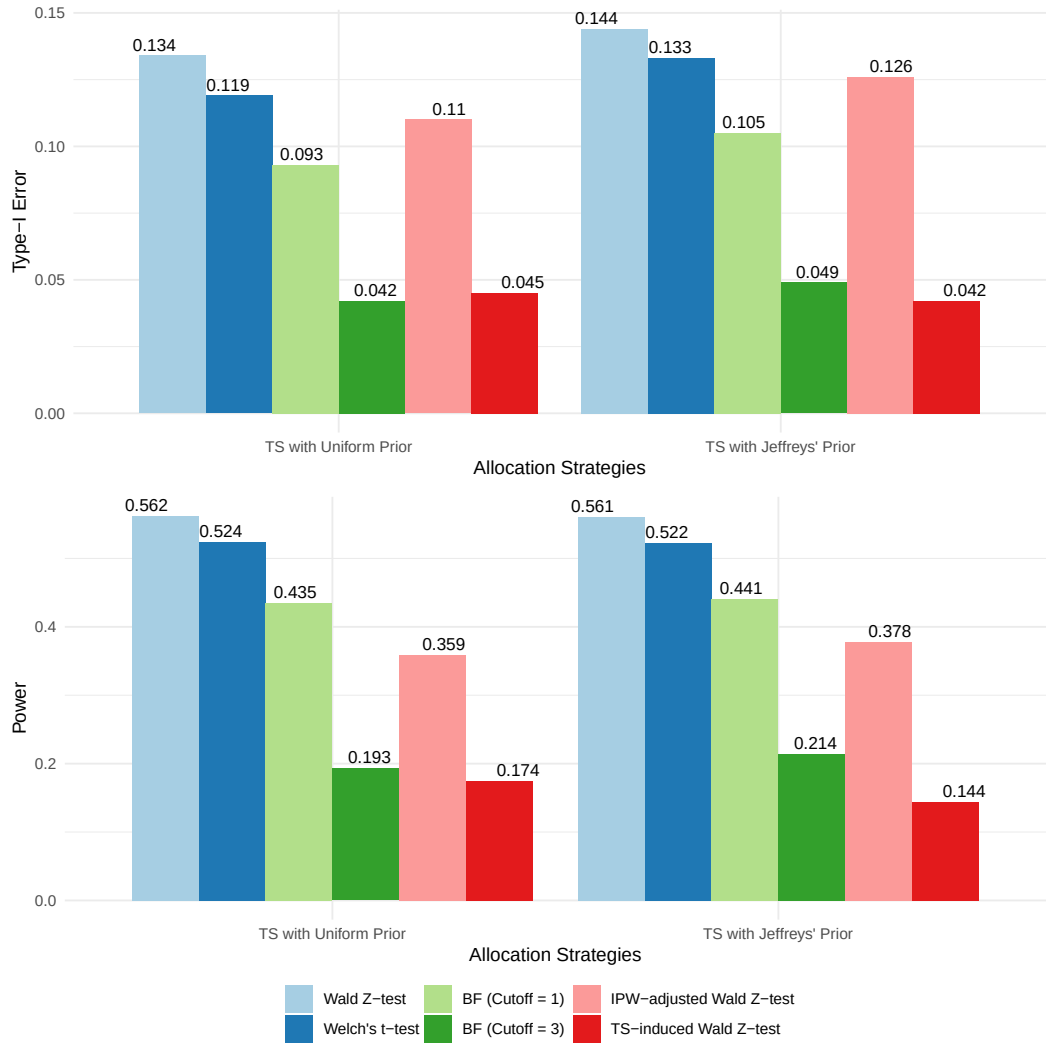

Figure E.1. Type-I error and power for data collected using Thompson Sampling (TS) with two different prior choices, i.e., Beta$(1, 1)$, equivalent to a Uniform distribution in $[0, 1]$, and the Jeffreys' prior, in this case a Beta$(1/2, 1/2)$. Hypothesis testing are based on a significance value $\alpha = 0.5$ and are performed with different statistical tests: 1. Wald Z-test, 2. Welch's t-test, 3. Bayes factor (BF) with cutoff 1, and 4. Bayes factor (BF) with cutoff 3. Type-I error and power are also reported for the adjustments of the Wald Z-test proposed in this work (IPW-adjusted Wald Z-test and TS-induced Wald Z-test). Results are based on a sample size of $n = 785$ and a number of independently simulated dataset of 5000.

F Sensitivity to Different Arm Means for TS-induced Wald-Z Test

As discussed in Section 4.4, the TS-induced Wald-Z test proposals requires the estimation of the empirical distribution of the Wald Z-test under the null hypothesis when data are collected with TS, in order to estimated the critical values able to control the type-I error. For the null hypothesis, the main text focuses on a case where $p_1 = p_2 = 0.50$. Here, we examine whether a different scenario would have an impact

on the distribution of the TS-induced Wald-Z test and other consequences on the type-I error and power, or whether this would be relatively constant across varying instantiations of the null hypothesis. We consider the case when $p_1 = p_2 = 0.25$.

Comparing hypothesis testing with data from $p_1 = p_2 = 0.5$ versus hypothesis testing with data from $p_1 = p_2 = 0.25$, Figure F.1 demonstrates that results are very similar: for Wald Z-test, Welch's t-test, the Bayes factor, and the IPW-adjusted Wald Z-test no type-I error value differ by more than 1%. For the TS-induced Wald-Z test, we also see relatively similar performance to the equivalent version in the main text: a type-I error of 5.2% for $p_1 = p_2 = 0.25$ compared to 4.5% for $p_1 = p_2 = 0.50$.

However, one should notice that in both cases, we have assumed that the values for the arm mean that is assumed by the algorithm-induced cutoff matches the true values of the arm mean. However, this information is not known to the experimenter ahead of time, and thus the TS-induced cutoffs may be set using a value that doesn't match the true rewards, even if some attempts are made to estimate the true values such as by taking the overall average reward across all samples. Thus, we also analysed the results of Wald's Z-test with TS-induced cutoffs when the cutoff simulations are based an incorrect arm mean; in particular, we chose the cutoffs based on an assumption that $p_1 = p_2 = 0.50$, but in actuality, $p_1 = p_2 = 0.25$. We found that type-I error is even better controlled, with a value of 5.0%, meaning that higher mean rewarsd will result in more conservative cutoffs for controlling type-I error, independently on the actual collected data.

G Non-linear Time Effect on the Steps-Change Variable

Several authors, including an external reviewer of this thesis, Ken Cheung, suggest that in mobile health application, time variables such as "Study Day" may have a non-linear effect on the outcome variable of interest.

We explored the non-linear effect of time by running GEE models with a polynomial relationship, first by adding a quadratic term for time (parabolic curve), and then an additional a cubic term as well. Our results also suggests this non-linear hypothesis, in which a quadratic effect of time resulted to be statistically significant. Here we report the results of our GEE models with a focus on the time variable only.

Covariate	Estimate	Standard Error	p-value
Study Day	1.20	2.17	0.58

Table G.8. Results of the GEE model studying the linear effect of time only on steps change. GEE: Generalized Estimating Equations.

As shown in the above tables, the linear effect of time seems to be positive, while the quadratic effect negative. Notably, with a second-order polynomial both the linear and quadratic effect were significant (positive and negative estimate, respectively), suggesting that at the beginning of the study there is an increase in the steps change with increasing time, but after a while, increasing time will result in a decrease in the steps change (negative coefficient estimate).

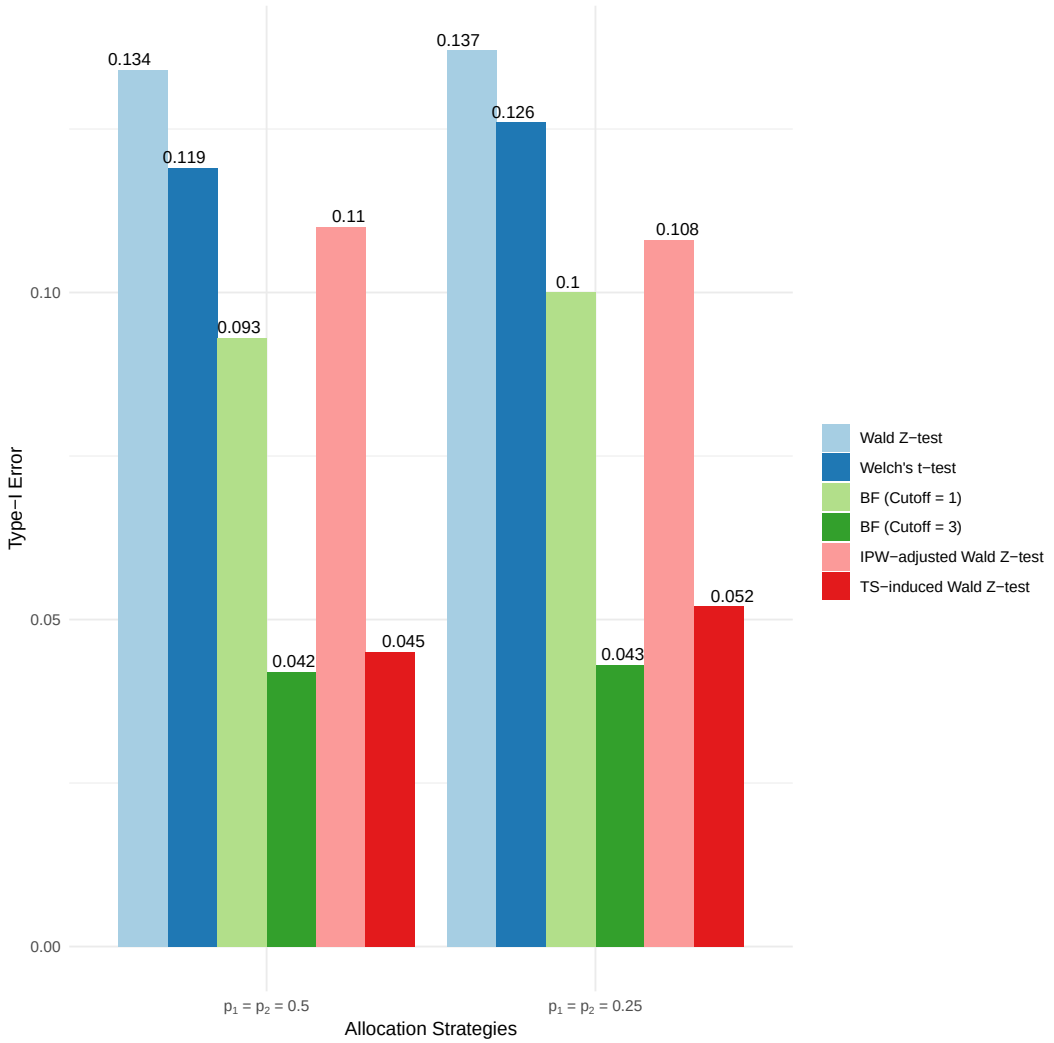

Figure F.1. Examining type-I error with two different null hypothesis scenarios, i.e, $p_1 = p_2 = 0.5$ and $p_1 = p_2 = 0.25$, with various hypothesis tests and the Thompson Sampling (TS) allocation strategy. Results are based on a sample size of $n = 785$ and a number of independently simulated dataset of 5000.

Covariate	Estimate	Standard Error	p-value
Study Day	24.53	9.94	0.01
I(Study Day, including also a quadratic effect of time. GEE: Generalized Estimating Equations.)	−0.51	0.20	0.01

Table G.8. Results of the GEE model studying the linear and quadratic effect of time on steps change. GEE: Generalized Estimating Equations.

Including both linear and quadratic term for time in the other multivariate models of interest did not change remarkably the final results, as shown in the tables below.

Covariate	Estimate	Standard Error	p-value
Study Day	66.81	37.96	0.07
I(Study Day^2)	−2.74	1.84	0.13
I(Study Day^3)	0.03	0.02	0.21

Table G.8. Results of the GEE model studying the linear, quadratic and cubic effect of time on steps change. GEE: Generalized Estimating Equations.

Covariate	Estimate	Standard Error	p-value
Message Sent	738.72	290.36	0.011
Study Day	55.82	17.06	0.001
I(Study Day^2)	−0.60	0.22	0.007
Message Sent*Study Day	−34.38	11.88	0.003

Table G.8. Results of the GEE model studying effects of sending a message on steps change, including also a quadratic effect of time. GEE: Generalized Estimating Equations.

Covariate	Estimate	Standard Error	p-value
Motivation	723.50	412.68	0.079
Feedback	−321.25	404.81	0.42
Study Day	42.45	15.32	0.005
I(Study Day^2)	−0.65	0.22	0.003
Motivation*Study Day	−15.41	15.39	0.316
Feedback*Study Day	−1.92	15.16	0.898
Motivation*Feedback	4.17	334.76	0.990

Table G.8. Results of the GEE model studying effects of motivational and feedback message on steps change, including also a quadratic effect of time. GEE: Generalized Estimating Equations.

Covariate	Estimate	Standard Error	p-value
M1. Capability/Self-belief	262	255.18	0.304
M2. Opportunity	417	234	0.075
M3. Motivation/Walk-benefit	417	242	0.084
F1. Reaching goal	−165	283	0.558
F2. Steps walked yesterday	−416	255	0.103
F3. Walked more/less than goal yesterday	−142	230	0.536
F4. Steps walked yesterday, plus a positive/negative message	−678	213	0.001
Study day	28.03	10.04	0.005
I(Study Day^2)	−0.57	0.20	0.005

Table G.8. Results of the GEE models studying the effects of different categories of feedback and motivation on steps change, including also a quadratic effect of time. GEE: Generalized Estimating Equations.

H Regression Analyses on the Uniform Random Group

As a sensitivity analysis for the text-messaging app study, we repeated all the analyses focusing only on the uniform random group ($n = 66$), for which stronger

statistical properties are demonstrated being a uniform and independent rule of assignment, thus not subject to bias due to adaptive nature that characterizes the Thompson Sampling. Results are reported in Tables H.8, H.8 and H.8.

Covariate	Estimate	95% CI	p-value
Message Sent	202	$[-274, 679]$	0.406
Study Day	10.5	$[-4.83, 25.8]$	0.180
Message*Study Day	-12.3	$[-31.1, 6.42]$	0.197

Table H.8. Results of the GEE model studying effects of sending any versus no message on steps change, based on the uniform random group only ($n = 66$). CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
Motivation	432	$[-301, 1164]$	0.248
Feedback	-510	$[-1224, 205]$	0.162
Study Day	2.33	$[-12.1, 16.8]$	0.752
Motivation*Study Day	-2.9	$[-29.2, 23.4]$	0.829
Feedback*Study Day	2.34	$[-23.8, 28.5]$	0.861
Motivation*Feedback	68.7	$[-535, 672]$	0.823

Table H.8. Results of the GEE model studying effects of motivational and feedback message on steps change, based on the uniform random group only ($n = 66$). CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
M1. Capability/Self-belief	295	$[-281, 871]$	0.315
M2. Opportunity	517	$[-24.6, 1060]$	0.06
M3. Motivation/Walk-benefit	441	$[-118, 1000]$	0.122
F1. Reaching goal	76.1	$[-554, 706]$	0.813
F2. Steps walked yesterday	-144	$[-714, 426]$	0.620
F3. Walked more/less than goal yesterday	-151	$[-653, 352]$	0.556
F4. Steps walked yesterday, plus a positive/negative message	-540	$[-1041, -40.7]$	0.034
Study day	-0.806	$[-5.42, 3.81]$	0.732

Table H.8. Results of the GEE models studying the effects of different categories of feedback and motivation on steps change, based on the uniform random group only ($n = 66$). CI: Confidence Interval, GEE: Generalized Estimating Equations.

Similarly to the main results, sending a message initially resulted in a positive effect on steps, but decreased over time. However, the effects were no longer significant which could be due to decreased power. The positive effect on steps seemed to be driven by motivational messages but this also lost significance after adding interaction terms. The main categories showing significance were again a self-efficacy message (a borderline $p = 0.07$ with an estimated coefficient of 517) and

a feedback message with the number of steps participants walked yesterday plus a negative/positive feedback message (-541 steps, $p = 0.03$).

I Multivariable Regression with Missing Data Imputation

We used multiple imputation as missing data imputation technique. More specifically, we employed multivariate imputation by chained equations, also called fully conditional specification or sequential regression multiple imputation. This method has emerged in the statistical literature as one principled method of addressing missing data, and a dedicated package in R exists. As we performed multiple imputation considering a number of three imputations, we report results for all the imputed datasets. This will allow assessment of the robustness of the method, as well a more exhaustive sensitivity analysis. Consistent with the main original findings, sensitivity results with data imputation show the same directions of effects.

Covariate	**Estimate**	**95% CI**	**p-value**
Message Sent	381	$[-1.82, 1044]$	0.050
Study Day	5.96	$[4.52, 41.3]$	0.014
Message*Study Day	-4.71	$[-44.9, -2.28]$	0.030

Table I.8. Results of the GEE model studying effects of sending any versus no message on steps change, with data from the 1st imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	**Estimate**	**95% CI**	**p-value**
Motivation	436	$[50.1, 1576]$	0.036
Feedback	-749	$[-1018, 394]$	0.387
Study Day	2.21	$[-4.04, 29.3]$	0.137
Motivation*Study Day	-1.06	$[-39, 12.2]$	0.304
Feedback*Study Day	0.02	$[-22.9, 26.6]$	0.882
Motivation*Feedback	63.5	$[-938, 396]$	0.426

Table I.8. Results of the GEE model studying effects of motivational and feedback message on steps change, with data from the 1st imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
M1. Capability/Self-belief	221	$[-269, 711]$	0.377
M2. Opportunity	482	$[56, 908]$	0.025
M3. Motivation/Walk-benefit	235	$[-215, 686]$	0.306
F1. Reaching goal	-308	$[-812, 195]$	0.230
F2. Steps walked yesterday	-450	$[-871, -30]$	0.035
F3. Walked more/less than goal yesterday	-226	$[-648, 197]$	0.296
F4. Steps walked yesterday, plus a positive/negative message	-572	$[-959, -185]$	0.003
Study day	6.38	$[-0.439, 13.2]$	0.066

Table I.8. Results of the GEE models studying the effects of different categories of feedback and motivation on steps change, with data from the 1st imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
Message Sent	84.3	$[-404, 572]$	0.735
Study Day	10.1	$[-7.8, 28]$	0.269
Message*Study Day	-9.37	$[-31.4, 12.7]$	0.404

Table I.8. Results of the GEE model studying effects of sending any versus no message on steps change, with data from the 2nd imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
Motivation	446	$[-370, 1063]$	0.343
Feedback	-449	$[-1167, 269]$	0.221
Study Day	3.16	$[-12.9, 19.2]$	0.699
Motivation*Study Day	0.14	$[-26.7, 27]$	0.992
Feedback*Study Day	0.99	$[-24.5, 26.5]$	0.939
Motivation*Feedback	-37	$[-9.37, 597]$	0.909

Table I.8. Results of the GEE model studying effects of motivational and feedback message on steps change, with data from the 2nd imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
M1. Capability/Self-belief	227	$[-186, 740]$	0.240
M2. Opportunity	381	$[-24.6, 787]$	0.065
M3. Motivation/Walk-benefit	287	$[-148, 723]$	0.196
F1. Reaching goal	-270	$[-759, 218]$	0.278
F2. Steps walked yesterday	-451	$[-887, -15.7]$	0.042
F3. Walked more/less than goal yesterday	-287	$[-696, 122]$	0.169
F4. Steps walked yesterday, plus a positive/negative message	-764	$[-1191, -336]$	< 0.001
Study day	3.85	$[-2.77, 10.5]$	0.255

Table I.8. Results of the GEE models studying the effects of different categories of feedback and motivation on steps change, with data from the 2nd imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
Message Sent	202	$[-274, 679]$	0.406
Study Day	10.5	$[-4.83, 25.8]$	0.180
Message*Study Day	-12.3	$[-31.1, 6.42]$	0.197

Table I.8. Results of the GEE model studying effects of sending any versus no message on steps change, with data from the 3rd imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
Motivation	432	$[-301, 1164]$	0.248
Feedback	-510	$[-1224, 205]$	0.162
Study Day	2.33	$[-12.1, 16.8]$	0.752
Motivation*Study Day	-2.90	$[-29.2, 23.4]$	0.829
Feedback*Study Day	2.34	$[-23.8, 28.5]$	0.861
Motivation*Feedback	68.7	$[-535, 672]$	0.823

Table I.8. Results of the GEE model studying effects of motivational and feedback message on steps change, with data from the 3rd imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Covariate	Estimate	95% CI	p-value
M1. Capability/Self-belief	249	$[-193, 691]$	0.269
M2. Opportunity	461	$[67.4, 855]$	0.021
M3. Motivation/Walk-benefit	473	$[20.5, 925]$	0.040
F1. Reaching goal	-128	$[-650, 393]$	0.630
F2. Steps walked yesterday	-451	$[-911, 9.44]$	0.054
F3. Walked more/less than goal yesterday	-250	$[-654, 154]$	0.224
F4. Steps walked yesterday, plus a positive/negative message	-839	$[-1228, -450]$	< 0.001
Study day	2.07	$[-4.01, 8.15]$	0.504

Table I.8. Results of the GEE models studying the effects of different categories of feedback and motivation on steps change, with data from the 3rd imputed dataset. CI: Confidence Interval, GEE: Generalized Estimating Equations.

Bibliography

Abbasi-Yadkori, Yasin, Pál, Dávid, & Szepesvári, Csaba. 2011. Improved algorithms for linear stochastic bandits. *Pages 2312–2320 of: Advances in neural information processing systems.*

Agrawal, Shipra, & Goyal, Navin. 2012. Analysis of thompson sampling for the multi-armed bandit problem. *Pages 39–1 of: Conference on learning theory.*

Agrawal, Shipra, & Goyal, Navin. 2013. Thompson sampling for contextual bandits with linear payoffs. *Pages 127–135 of: International conference on machine learning.*

Agresti, Alan. 2003. *Categorical data analysis.* Vol. 482. John Wiley & Sons.

Aguilera, Adrian, Figueroa, Caroline A, Hernandez-Ramos, Rosa, Sarkar, Urmimala, Cemballi, Anupama, Gomez-Pathak, Laura, Miramontes, Jose, Yom-Tov, Elad, Chakraborty, Bibhas, Yan, Xiaoxi, *et al.* 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.

Ahn, Inkyung, & Park, Jooyoung. 2011. Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *Biosystems*, **106**(2-3), 121–129.

Ahuja, Vishal, & Birge, John R. 2016. Response-adaptive designs for clinical trials: Simultaneous learning from multiple patients. *European journal of operational research*, **248**(2), 619–633.

Ahuja, Vishal, & Birge, John R. 2020. An approximation approach for response-adaptive clinical trial design. *Informs journal on computing*, **32**(4), 877–894.

Akrour, Riad, Schoenauer, Marc, & Sebag, Michèle. 2012. April: Active preference learning-based reinforcement learning. *Pages 116–131 of: Joint european conference on machine learning and knowledge discovery in databases.* Springer.

Albright, Ann L, & Gregg, Edward W. 2013. Preventing type 2 diabetes in communities across the us: the national diabetes prevention program. *American journal of preventive medicine*, **44**(4), S346–S351.

Allen, Luke Nelson, & Christie, Gillian Pepall. 2016. The emergence of personalized health technology. *Journal of medical internet research*, **18**(5), e99.

Almirall, Daniel, Nahum-Shani, Inbal, Sherwood, Nancy E, & Murphy, Susan A. 2014. Introduction to smart designs for the development of adaptive interventions: with application to weight loss research. *Translational behavioral medicine*, **4**(3), 260–274.

Antognini, Alessandro Baldi, & Giovagnoli, Alessandra. 2015. *Adaptive designs for sequential treatment allocation*. Vol. 73. CRC Press.

Arjas, Elja, & Saarela, Olli. 2010. Optimal dynamic regimes: presenting a case for predictive inference. *The international journal of biostatistics*, **6**(2).

Asoh, Hideki, Shiro, Masanori, Akaho, Shotaro, Kamishima, Toshihiro, Hashida, K, Aramaki, Eiji, & Kohro, Takahide. 2013. Modeling medical records of diabetes using markov decision processes. *In: Proceedings of icml2013 workshop on role of machine learning in transforming healthcare*.

Aswani, Anil, Kaminsky, Philip, Mintz, Yonatan, Flowers, Elena, & Fukuoka, Yoshimi. 2019. Behavioral modeling in weight loss interventions. *European journal of operational research*, **272**(3), 1058–1072.

Atan, Onur, Jordon, James, & van der Schaar, Mihaela. 2018. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. *Pages 2071–2078 of: Aaai*.

Atan, Onur, Zame, William R, & Schaar, Mihaela. 2019. Sequential patient recruitment and allocation for adaptive clinical trials. *Pages 1891–1900 of: The 22nd international conference on artificial intelligence and statistics*. PMLR.

Atas, Amine Hatun, & Çelik, Berkan. 2019. Smartphone use of university students: Patterns, purposes, and situations. *Malaysian online journal of educational technology*, **7**(2), 59–70.

Athreya, Krishna B, & Karlin, Samuel. 1968. Embedding of urn schemes into continuous time markov branching processes and related limit theorems. *The annals of mathematical statistics*, **39**(6), 1801–1817.

Atkeson, Christopher G., & Santamaria, Juan Carlos. 1997. A comparison of direct and model-based reinforcement learning. *Pages 3557–3564 of: In international conference on robotics and automation*. IEEE Press.

Atkinson, Anthony C, *et al.* 2014. Selecting a biased-coin design. *Statistical science*, **29**(1), 144–163.

Audibert, Jean-Yves, & Bubeck, Sébastien. 2010. Best arm identification in multi-armed bandits. *In: 23rd Conference on Learning Theory (COLT)*.

Auer, Peter. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of machine learning research*, **3**(Nov), 397–422.

Auer, Peter, Cesa-Bianchi, Nicolo, & Fischer, Paul. 2002a. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, **47**(2-3), 235–256.

Auer, Peter, Cesa-Bianchi, Nicolo, Freund, Yoav, & Schapire, Robert E. 2002b. The nonstochastic multiarmed bandit problem. *Siam journal on computing*, **32**(1), 48–77.

Avila-Garcia, Patricia, Hernandez-Ramos, Rosa, Nouri, Sarah S, Cemballi, Anupama, Sarkar, Urmimala, Lyles, Courtney R, & Aguilera, Adrian. 2019. Engaging users in the design of an mhealth, text message-based intervention to increase physical activity at a safety-net health care system. *Jamia open*, **2**(4), 489–497.

Aziz, Maryam, Kaufmann, Emilie, & Riviere, Marie-Karelle. 2019. On multi-armed bandit designs for phase i clinical trials. *Arxiv*, **abs/1903.07082**.

Azur, Melissa J, Stuart, Elizabeth A, Frangakis, Constantine, & Leaf, Philip J. 2011. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, **20**(1), 40–49.

Bakshy, Eytan, Eckles, Dean, Yan, Rong, & Rosenn, Itamar. 2012. Social influence in social advertising: evidence from field experiments. *Pages 146–161 of: Proceedings of the 13th acm conference on electronic commerce*.

Bartroff, Jay, & Lai, Tze Leung. 2010. Approximate dynamic programming and its applications to the design of phase i cancer trials. *Statistical science*, **25**(2), 245–257.

Bartroff, Jay, Lai, Tze Leung, & Shih, Mei-Chiung. 2013. *Sequential ex-perimentation in clinical trials.* Springer.

Bates, Douglas, Mächler, Martin, Bolker, Ben, & Walker, Steve. 2014. Fitting linear mixed-effects models using lme4. *arxiv preprint arxiv:1406.5823*.

Bather, John. 2000. *Decision theory: An introduction to dynamic programming and sequential decisions*. John Wiley & Sons, Inc.

Baumel, Amit, Muench, Frederick, Edan, Stav, & Kane, John M. 2019. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. *Journal of medical internet research*, **21**(9), e14567.

Bell, Lauren, Garnett, Claire, Qian, Tianchen, Perski, Olga, Potts, Henry WW, & Williamson, Elizabeth. 2020. Notifications to improve engagement with an alcohol reduction app: Protocol for a micro-randomized trial. *Jmir research protocols*, **9**(8), e18690.

Bellman, Richard. 1956. A problem in the sequential design of experiments. *Sankhyā: The indian journal of statistics (1933-1960)*, **16**(3/4), 221–229.

Bellman, Richard. 1957. *Dynamic programming*. 1 edn. Princeton, NJ, USA: Princeton University Press.

Ben-Zeev, Dror, Brenner, Christopher J, Begale, Mark, Duffecy, Jennifer, Mohr, David C, & Mueser, Kim T. 2014. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia bulletin*, **40**(6), 1244–1253.

Bernell, Stephanie, & Howard, Steven W. 2016. Use your words carefully: what is a chronic disease? *Frontiers in public health*, **4**, 159.

Berry, D. A., & Fristedt, B. 1985a. *Bandit problems: Sequential allocation of experiments*. Chapman and Hall, London.

Berry, Donald A. 2006. Bayesian clinical trials. *Nature reviews drug discovery*, **5**(1), 27–36.

Berry, Donald A, & Fristedt, Bert. 1985b. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and hall*, **5**(71-87), 7–7.

Berry, Scott M, Carlin, Bradley P, Lee, J Jack, & Muller, Peter. 2010. *Bayesian adaptive methods for clinical trials*. CRC press.

Bertsekas, Dimitri P. 2011. Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, ma: Athena scientific*.

Bertsekas, Dimitri P. 2019. *Reinforcement learning and optimal control*. Athena Scientific Belmont, MA.

Bertsekas, Dimitri P, & Tsitsiklis, John N. 1996. *Neuro-dynamic programming*. Athena Scientific.

Bhatt, Deepak L, & Mehta, Cyrus. 2016. Adaptive designs for clinical trials. *New england journal of medicine*, **375**(1), 65–74.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. springer.

Blatt, Doron, Murphy, Susan A, & Zhu, Ji. 2004. A-learning for approximate planning. *Ann arbor*, **1001**, 48109–2122.

Bolger, Niall, & Laurenceau, Jean-Philippe. 2013. *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Boruvka, Audrey, Almirall, Daniel, Witkiewitz, Katie, & Murphy, Susan A. 2018. Assessing time-varying causal effect moderation in mobile health. *Journal of the american statistical association*, **113**(523), 1112–1121.

Bothwell, Laura E, Greene, Jeremy A, Podolsky, Scott H, Jones, David S, *et al.* 2016. Assessing the gold standard—lessons from the history of rcts. *N engl j med*, **374**(22), 2175–2181.

Bouton, Mark E. 2007. *Learning and behavior: A contemporary synthesis*. Sinauer Associates.

Bowden, Jack, & Trippa, Lorenzo. 2017. Unbiased estimation for response adaptive clinical trials. *Statistical methods in medical research*, **26**(5), 2376–2388.

Breiman, Leo, *et al.* 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, **16**(3), 199–231.

Brugnara, Laura, Murillo, Serafín, Novials, Anna, Rojo-Martínez, Gemma, Soriguer, Federico, Goday, Albert, Calle-Pascual, Alfonso, Castaño, Luis, Gaztambide, Sonia, Valdés, Sergio, *et al.* 2016. Low physical activity and its association with diabetes and other cardiovascular risk factors: a nationwide, population-based study. *Plos one*, **11**(8), e0160959.

Bubeck, Sébastien, & Cesa-Bianchi, Nicolo. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arxiv preprint arxiv:1204.5721*.

Bubeck, Sébastien, Munos, Rémi, & Stoltz, Gilles. 2009. Pure exploration in multi-armed bandits problems. *Pages 23–37 of: International conference on algorithmic learning theory*. Springer.

Bubeck, Sébastien, Munos, Rémi, & Stoltz, Gilles. 2011. Pure exploration in finitely-armed and continuous-armed bandits. *Theor. comput. sci.*, **412**, 1832–1852.

Burnett, Thomas, Mozgunov, Pavel, Pallmann, Philip, Villar, Sofia S, Wheeler, Graham M, & Jaki, Thomas. 2020. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. *Bmc medicine*, **18**(1), 1–21.

Busa-Fekete, Róbert, Szörényi, Balázs, Weng, Paul, Cheng, Weiwei, & Hüllermeier, Eyke. 2014. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine learning*, **97**(3), 327–351.

Butler, Emily L, Laber, Eric B, Davis, Sonia M, & Kosorok, Michael R. 2018. Incorporating patient preferences into estimation of optimal individualized treatment rules. *Biometrics*, **74**(1), 18–26.

Camerer, Colin F, Dreber, Anna, Forsell, Eskil, Ho, Teck-Hua, Huber, Jürgen, Johannesson, Magnus, Kirchler, Michael, Almenberg, Johan, Altmejd, Adam, Chan, Taizan, *et al.* 2016. Evaluating replicability of laboratory experiments in economics. *Science*, **351**(6280), 1433–1436.

Carney, Rebekah, Bradshaw, Tim, & Yung, Alison R. 2016. Physical health promotion for young people at ultra-high risk for psychosis: An application of the com-b model and behaviour-change wheel. *International journal of mental health nursing*, **25**(6), 536–545.

Casella, George, & Berger, Roger L. 2002. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.

Castro, Oscar, Bennie, Jason, Vergeer, Ineke, Bosselut, Grégoire, & Biddle, Stuart JH. 2020. How sedentary are university students? a systematic review and meta-analysis. *Prevention science*, **21**(3), 332–343.

CDC. 2020. Prevalence of Self-Reported Physical Inactivity Among US Adults by Race/Ethnicity, State and Territory, BRFSS, 2015–2018. . *Center for disease control and prevention.*

Cella, Leonardo, & Cesa-Bianchi, Nicolò. 2020. Stochastic bandits with delay-dependent payoffs. *Pages 1168–1177 of: International conference on artificial intelligence and statistics.*

Chakraborty, Bibhas, & Moodie, Erica E.M. 2013. *Statistical methods for dynamic treatment regimes: Reinforcement learning, causal inference, and personalized medicine.* Springer.

Chakraborty, Bibhas, & Murphy, Susan A. 2014. Dynamic treatment regimes. *Annual review of statistics and its application*, **1**, 447–464.

Chakraborty, Bibhas, Strecher, Victor, & Murphy, SA. 2008. Bias correction and confidence intervals for fitted q-iteration. *In: Workshop on model uncertainty and risk in reinforcement learning, nips, whistler, canada.* Citeseer.

Chakraborty, Bibhas, Murphy, Susan, & Strecher, Victor. 2010. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical methods in medical research*, **19**(3), 317–343.

Chakraborty, Bibhas, Laber, Eric B, & Zhao, Yingqi. 2013. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, **69**(3), 714–723.

Chakravorty, Jhelum, & Mahajan, Aditya. 2014. Multi-armed bandits, gittins index, and its calculation. *Methods and applications of statistics in clinical trials: Planning, analysis, and inferential methods*, **2**(416-435), 455.

Chapelle, Olivier, & Li, Lihong. 2011. An empirical evaluation of thompson sampling. *Pages 2249–2257 of: Advances in neural information processing systems.*

Chen, Jingxiang, Fu, Haoda, He, Xuanyao, Kosorok, Michael R, & Liu, Yufeng. 2018. Estimating individualized treatment rules for ordinal treatments. *Biometrics*, **74**(3), 924–933.

Chen, Ming-Hui, Müller, Peter, Sun, Dongchu, Ye, Keying, & Dey, Dipak K. 2010. *Frontiers of statistical decision making and bayesian analysis: In honor of james o. berger.* Springer Science & Business Media.

Cheng, Weiwei, Fürnkranz, Johannes, Hüllermeier, Eyke, & Park, Sang-Hyeun. 2011. Preference-based policy iteration: Leveraging preference learning for reinforcement learning. *Pages 312–327 of: Joint european conference on machine learning and knowledge discovery in databases.* Springer.

Cheung, Ying Kuen. 2015. A review of dose finding methods and theory. *Communications for statistical applications and methods*, **22**(5), 401–413.

Cheung, Ying Kuen, Inoue, Lurdes YT, Wathen, J Kyle, & Thall, Peter F. 2006. Continuous bayesian adaptive randomization based on event times with covariates. *Statistics in medicine*, **25**(1), 55–70.

Cheung, Ying Kuen, Chakraborty, Bibhas, & Davidson, Karina W. 2015. Sequential multiple assignment randomized trial (smart) with adaptive randomization for quality improvement in depression treatment program. *Biometrics*, **71**(2), 450–459.

Chevret, S. 2006. *Statistical methods for dose-finding experiments*. Statistics in Practice. Wiley.

Choi, Karmel W, Chen, Chia-Yen, Stein, Murray B, Klimentidis, Yann C, Wang, Min-Jung, Koenen, Karestan C, & Smoller, Jordan W. 2019. Assessment of bidirectional relationships between physical activity and depression among adults: a 2-sample mendelian randomization study. *Jama psychiatry*, **76**(4), 399–408.

Chow, Shein-Chung, & Chang, Mark. 2008. Adaptive design methods in clinical trials–a review. *Orphanet journal of rare diseases*, **3**(1), 11.

Chow, Shein-Chung, Chang, Mark, & Pong, Annpey. 2005. Statistical consideration of adaptive methods in clinical development. *Journal of biopharmaceutical statistics*, **15**(4), 575–591.

Chu, Wei, Li, Lihong, Reyzin, Lev, & Schapire, Robert. 2011. Contextual bandits with linear payoff functions. *Pages 208–214 of: Proceedings of the fourteenth international conference on artificial intelligence and statistics*.

Clement, Benjamin, Roy, Didier, Oudeyer, Pierre-Yves, & Lopes, Manuel. 2014. Online optimization of teaching sequences with multi-armed bandits. *In: 7th international conference on educational data mining*.

Clement, Benjamin, Roy, Didier, Oudeyer, Pierre-Yves, & Lopes, Manuel. 2015. Multi-Armed Bandits for Intelligent Tutoring Systems. *Journal of educational data mining*, **7**(2), 20–48. The file is in PDF format. If your computer does not recognize it, simply download the file and then open it with your browser.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences, 2nd edn. á/l*.

Collaboration, Open Science, *et al.* 2015. Estimating the reproducibility of psychological science. *Science*, **349**(6251).

Collins, Linda M, Murphy, Susan A, & Bierman, Karen L. 2004. A conceptual framework for adaptive preventive interventions. *Prevention science*, **5**(3), 185–196.

Collins, Linda M, Murphy, Susan A, Nair, Vijay N, & Strecher, Victor J. 2005. A strategy for optimizing and evaluating behavioral interventions. *Annals of behavioral medicine*, **30**(1), 65–73.

Collins, Linda M, Chakraborty, Bibhas, Murphy, Susan A, & Strecher, Victor. 2009. Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. *Clinical trials*, **6**(1), 5–15.

Collins, LM, Murphy, SA, & Bierman, KA. 2001. Design, and evaluation of adaptive preventive interventions. *Prevention science*.

Consolvo, Sunny, McDonald, David W, Toscos, Tammy, Chen, Mike Y, Froehlich, Jon, Harrison, Beverly, Klasnja, Predrag, LaMarca, Anthony, LeGrand, Louis, Libby, Ryan, *et al.* 2008. Activity sensing in the wild: a field trial of ubifit garden. *Pages 1797–1806 of: Proceedings of the sigchi conference on human factors in computing systems*.

Cronbach, Lee J. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, **16**(3), 297–334.

Crump, Richard K, Hotz, V Joseph, Imbens, Guido W, & Mitnik, Oscar A. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, **96**(1), 187–199.

Curran, Geoffrey M, Bauer, Mark, Mittman, Brian, Pyne, Jeffrey M, & Stetler, Cheryl. 2012. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Medical care*, **50**(3), 217.

Dallery, Jesse, & Raiff, Bethany R. 2014. Optimizing behavioral health interventions with single-case designs: from development to dissemination. *Translational behavioral medicine*, **4**(3), 290–303.

Dallery, Jesse, Cassidy, Rachel N, & Raiff, Bethany R. 2013. Single-case experimental designs to evaluate novel technology-based health interventions. *Journal of medical internet research*, **15**(2), e22.

Daniel, Rhian M, Cousens, SN, De Stavola, BL, Kenward, Michael G, & Sterne, JAC. 2013. Methods for dealing with time-dependent confounding. *Statistics in medicine*, **32**(9), 1584–1618.

Dash, Sabyasachi, Shakyawar, Sushil Kumar, Sharma, Mohit, & Kaushik, Sandeep. 2019. Big data in healthcare: management, analysis and future prospects. *Journal of big data*, **6**(1), 54.

Daskalaki, Elena, Diem, Peter, & Mougiakakou, Stavroula G. 2013. An actor–critic based controller for glucose regulation in type 1 diabetes. *Computer methods and programs in biomedicine*, **109**(2), 116–125.

Daskalova, Nediyana, Yoon, Jina, Wang, Yibing, Araujo, Cintia, Beltran Jr, Guillermo, Nugent, Nicole, McGeary, John, Williams, Joseph Jay, & Huang, Jeff. 2020. Sleepbandits: Guided flexible self-experiments for sleep. *Pages 1–13 of: Proceedings of the 2020 chi conference on human factors in computing systems*.

Dawson, Ree, & Lavori, Philip W. 2012. Efficient design and inference for multistage randomized trials of individualized treatment policies. *Biostatistics*, **13**(1), 142–152.

Deo, Rahul C. 2015. Machine learning in medicine. *Circulation*, **132**(20), 1920–1930.

Derrick, Ben, Toher, Deirdre, & White, Paul. 2016. Why welch's test is type i error robust. *The quantitative methods in psychology*, **12**(1).

Deshpande, Yash, Mackey, Lester, Syrgkanis, Vasilis, & Taddy, Matt. 2018. Accurate inference for adaptive linear models. *Pages 1194–1203 of: International conference on machine learning*. PMLR.

Dimairo, Munyaradzi, Pallmann, Philip, Wason, James, Todd, Susan, Jaki, Thomas, Julious, Steven A, Mander, Adrian P, Weir, Christopher J, Koenig, Franz, Walton, Marc K, *et al.* 2020. The adaptive designs consort extension (ace) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. *bmj*, **369**.

Dimakopoulou, Maria, Zhou, Zhengyuan, Athey, Susan, & Imbens, Guido. 2019. Balanced linear contextual bandits. *Pages 3445–3453 of: Proceedings of the aaai conference on artificial intelligence*, vol. 33.

Doya, Kenji. 2000. Reinforcement learning in continuous time and space. *Neural computation*, **12**(1), 219–245.

Duncan, Markus J, Wunderlich, Kelly, Zhao, Yingying, & Faulkner, Guy. 2018. Walk this way: validity evidence of iphone health application step count in laboratory and free-living conditions. *Journal of sports sciences*, **36**(15), 1695–1704.

Durham, SD, & Yu, KF. 1990. Randomized play-the-leader rules for sequential sampling from two populations. *Probability in the engineering and informational sciences*, **4**(3), 355–367.

Durham, SD, Flournoy, N, & Li, W. 1998. A sequential design for maximizing the probability of a favourable response. *Canadian journal of statistics*, **26**(3), 479–495.

Eckles, Dean, & Kaptein, Maurits. 2014. Thompson sampling with the online bootstrap. *arxiv preprint arxiv:1410.4009*.

Eckles, Dean, & Kaptein, Maurits. 2019. Bootstrap thompson sampling and sequential decision problems in the behavioral sciences. *Sage open*, **9**(2), 2158244019851675.

Efron, Bradley. 1971a. Forcing a sequential experiment to be balanced. *Biometrika*, **58**(3), 403–417.

Efron, Bradley. 1971b. Forcing a sequential experiment to be balanced. *Biometrika*, **58**(3), 403–417.

Efron, Bradley. 2012. Bayesian inference and the parametric bootstrap. *The annals of applied statistics*, **6**(4), 1971.

Eisele, Jeffrey R. 1994. The doubly adaptive biased coin design for sequential clinical trials. *Journal of statistical planning and inference*, **38**(2), 249–261.

Eisele, Jeffrey R, & Woodroofe, Michael B. 1995. Central limit theorems for doubly adaptive biased coin designs. *The annals of statistics*, 234–254.

Elliot, Andrew J, & Church, Marcy A. 1997. A hierarchical model of approach and avoidance achievement motivation. *Journal of personality and social psychology*, **72**(1), 218.

Engle, Robert F. 1984. Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, **2**, 775–826.

Epstein, Leonard H, Temple, Jennifer L, Roemmich, James N, & Bouton, Mark E. 2009. Habituation as a determinant of human food intake. *Psychological review*, **116**(2), 384.

Ernst, Damien, Stan, Guy-Bart, Goncalves, Jorge, & Wehenkel, Louis. 2006. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. *Pages 667–672 of: Proceedings of the 45th ieee conference on decision and control*. IEEE.

Erraqabi, Akram, Lazaric, Alessandro, Valko, Michal, Brunskill, Emma, & Liu, Yun-En. 2017. Trading off rewards and errors in multi-armed bandits. *Pages 709–717 of: Artificial intelligence and statistics*.

Ertefaie, Ashkan, & Strawderman, Robert L. 2018. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, **105**(4), 963–977.

Ertefaie, Ashkan, Shortreed, Susan, & Chakraborty, Bibhas. 2016. Q-learning residual analysis: application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia. *Statistics in medicine*, **35**(13), 2221–2234.

Escandell-Montero, Pablo, Martínez-Martínez, José M, Martín-Guerrero, José D, Soria-Olivas, Emilio, Vila-Francés, Joan, & Magdalena-Benedito, Rafael. 2011. Adaptive treatment of anemia on hemodialysis patients: A reinforcement learning approach. *Pages 44–49 of: 2011 ieee symposium on computational intelligence and data mining (cidm)*. IEEE.

Even-Dar, Eyal, Mannor, Shie, & Mansour, Yishay. 2002. Pac bounds for multi-armed bandit and markov decision processes. *Pages 255–270 of: International Conference on Computational Learning Theory*. Springer.

Even-Dar, Eyal, Mannor, Shie, & Mansour, Yishay. 2006. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, **7**(Jun), 1079–1105.

Eysenbach, Gunther. 2006. The law of attrition revisited–author's reply. *Journal of medical internet research*, **8**(3), 73–74.

Fan, Yanqin, He, Ming, Su, Liangjun, & Zhou, Xiao-Hua. 2019. A smoothed q-learning algorithm for estimating optimal dynamic treatment regimes. *Scandinavian journal of statistics*, **46**(2), 446–469.

FDA. 2019. Adaptive designs for clinical trials of drugs and biologics: Guidance for industry. *Washington dc, usa: Food and drug administration.* November 2019 (accessed June, 2020).

Figueroa, Caroline A, DeMasi, Orianna, Hernandez-Ramos, Rosa, & Aguilera, Adrian. 2020. Who benefits most from adding technology to depression treatment and how? an analysis of engagement with a texting adjunct for psychotherapy. *Telemedicine and e-health.*

Filippi, Sarah, Cappe, Olivier, Garivier, Aurélien, & Szepesvári, Csaba. 2010. Parametric bandits: The generalized linear case. *Pages 586–594 of: Advances in neural information processing systems.*

Firth, Joseph, Torous, John, Nicholas, Jennifer, Carney, Rebekah, Rosenbaum, Simon, & Sarris, Jerome. 2017a. Can smartphone mental health interventions reduce symptoms of anxiety? a meta-analysis of randomized controlled trials. *Journal of affective disorders*, **218**, 15–22.

Firth, Joseph, Torous, John, Nicholas, Jennifer, Carney, Rebekah, Pratap, Abhishek, Rosenbaum, Simon, & Sarris, Jerome. 2017b. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World psychiatry*, **16**(3), 287–298.

Fisher, Ronald Aylmer. 1992. Statistical methods for research workers. *Pages 66–70 of: Breakthroughs in statistics.* Springer.

Fitzmaurice, Garrett M, Laird, Nan M, & Ware, James H. 2012. *Applied longitudinal analysis.* Vol. 998. John Wiley & Sons.

Flannery, Caragh, McHugh, Sheena, Anaba, Ann Ebere, Clifford, E, O'Riordan, Mairead, Kenny, Louise C, McAuliffe, Fionnuala M, Kearney, Patricia M, & Byrne, M. 2018. Enablers and barriers to physical activity in overweight and obese pregnant women: an analysis informed by the theoretical domains framework and com-b model. *Bmc pregnancy and childbirth*, **18**(1), 178.

Forman, Evan M, Kerrigan, Stephanie G, Butryn, Meghan L, Juarascio, Adrienne S, Manasse, Stephanie M, Ontañón, Santiago, Dallal, Diane H, Crochiere, Rebecca J, & Moskow, Danielle. 2019. Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of behavioral medicine*, **42**(2), 276–290.

Fraley, R. Chris, & Vazire, Simine. 2014. The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *Plos one*, **9**(10), 1–12.

Friedman, Lawrence M, Furberg, Curt D, DeMets, David L, Reboussin, David M, & Granger, Christopher B. 2015. *Fundamentals of clinical trials*. Springer.

Friston, Karl. 2012. Ten ironic rules for non-statistical reviewers. *Neuroimage*, **61**(4), 1300 – 1310.

Fu, Sheng, He, Qinying, Zhang, Sanguo, & Liu, Yufeng. 2019. Robust outcome weighted learning for optimal individualized treatment rules. *Journal of biopharmaceutical statistics*, **29**(4), 606–624.

Fukuoka, Yoshimi, Haskell, William, Lin, Feng, & Vittinghoff, Eric. 2019. Short-and long-term effects of a mobile phone app in conjunction with brief in-person counseling on physical activity among physically inactive women: the mped randomized clinical trial. *Jama network open*, **2**(5), e194281–e194281.

Fürnkranz, Johannes, Hüllermeier, Eyke, Cheng, Weiwei, & Park, Sang-Hyeun. 2012. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, **89**(1-2), 123–156.

Gal, Roxanne, May, Anne M, van Overmeeren, Elon J, Simons, Monique, & Monninkhof, Evelyn M. 2018. The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. *Sports medicine-open*, **4**(1), 1–15.

Garnett, Claire, Crane, David, West, Robert, Brown, Jamie, & Michie, Susan. 2019. The development of drink less: an alcohol reduction smartphone app for excessive drinkers. *Translational behavioral medicine*, **9**(2), 296–307.

Geurts, Pierre, Ernst, Damien, & Wehenkel, Louis. 2006. Extremely randomized trees. *Machine learning*, **63**(1), 3–42.

Gittins, John. 1974. A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, 241–266.

Gittins, John, Glazebrook, Kevin, & Weber, Richard. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

Goldberg, Yair, & Kosorok, Michael R. 2012. Q-learning with censored data. *Annals of statistics*, **40**(1), 529.

Goldberg, Yair, Song, Rui, Kosorok, Michael R, *et al.* 2013. Adaptive q-learning. *Pages 150–162 of: From probability to statistics and back: High-dimensional models and processes–a festschrift in honor of jon a. wellner*. Institute of Mathematical Statistics.

Goldstein, Stephanie P, Evans, Brittney C, Flack, Daniel, Juarascio, Adrienne, Manasse, Stephanie, Zhang, Fengqing, & Forman, Evan M. 2017. Return of the jitai: applying a just-in-time adaptive intervention framework to the development of m-health solutions for addictive behaviors. *International journal of behavioral medicine*, **24**(5), 673–682.

Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron, & Bengio, Yoshua. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

Gottesman, Omer, Johansson, Fredrik, Komorowski, Matthieu, Faisal, Aldo, Sontag, David, Doshi-Velez, Finale, & Celi, Leo Anthony. 2019. Guidelines for reinforcement learning in healthcare. *Nat med*, **25**(1), 16–18.

Greenewald, Kristjan, Tewari, Ambuj, Murphy, Susan, & Klasnja, Predag. 2017. Action centered contextual bandits. *Pages 5977–5985 of: Advances in neural information processing systems*.

Grondman, Ivo, Busoniu, Lucian, Lopes, Gabriel AD, & Babuska, Robert. 2012. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *Ieee transactions on systems, man, and cybernetics, part c (applications and reviews)*, **42**(6), 1291–1307.

Guan, Yang, Li, Shengbo Eben, Duan, Jingliang, Li, Jie, Ren, Yangang, & Cheng, Bo. 2019. Direct and indirect reinforcement learning. *arxiv preprint arxiv:1912.10600*.

Gupta, Neha, Granmo, Ole-Christoffer, & Agrawala, Ashok. 2011. Thompson sampling for dynamic multi-armed bandits. *Pages 484–489 of: 2011 10th international conference on machine learning and applications and workshops*, vol. 1. IEEE.

Gustafson, David H, McTavish, Fiona M, Chih, Ming-Yuan, Atwood, Amy K, Johnson, Roberta A, Boyle, Michael G, Levy, Michael S, Driscoll, Hilary, Chisholm, Steven M, Dillenburg, Lisa, *et al.* 2014. A smartphone application to support recovery from alcoholism: a randomized clinical trial. *Jama psychiatry*, **71**(5), 566–572.

Hadad, Vitor, Hirshberg, David A, Zhan, Ruohan, Wager, Stefan, & Athey, Susan. 2019. Confidence intervals for policy evaluation in adaptive experiments. *arxiv preprint arxiv:1911.02768*.

Halekoh, Ulrich, Højsgaard, Søren, Yan, Jun, *et al.* 2006. The r package geepack for generalized estimating equations. *Journal of statistical software*, **15**(2), 1–11.

Hamer, Robert M, & Simpson, Pippa M. 2009. *Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials*.

Hardeman, Wendy, Houghton, Julie, Lane, Kathleen, Jones, Andy, & Naughton, Felix. 2019. A systematic review of just-in-time adaptive interventions (jitais) to promote physical activity. *International journal of behavioral nutrition and physical activity*, **16**(1), 31.

Harries, Tim, Eslambolchilar, Parisa, Stride, Chris, Rettie, Ruth, & Walton, Simon. 2013. Walking in the wild–using an always-on smartphone application to increase physical activity. *Pages 19–36 of: Ifip conference on human-computer interaction*. Springer.

Hassani, Amin, *et al.* 2010. Reinforcement learning based control of tumor growth with chemotherapy. *Pages 185–189 of: 2010 international conference on system science and engineering.* IEEE.

Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2009. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hauser, John R, Urban, Glen L, Liberali, Guilherme, & Braun, Michael. 2009. Website morphing. *Marketing science*, **28**(2), 202–223.

He, Jianxing, Baxter, Sally L, Xu, Jie, Xu, Jiming, Zhou, Xingtao, & Zhang, Kang. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, **25**(1), 30–36.

Hoel, David, & Sobel, Milton. 1971. *New sequential procedures for selecting the best of k binomial populations, with tables and comparisons.* Tech. rept. University of Minnesota.

Hoffman, Matthew, Shahriari, Bobak, & Freitas, Nando. 2014. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. *Pages 365–374 of: Artificial intelligence and statistics.* PMLR.

Hogg, Robert V, Tanis, Elliot A, & Zimmerman, Dale L. 1977. *Probability and statistical inference.* Vol. 993. Macmillan New York.

Holloway, S. T., Laber, E. B., Linn, K. A., Zhang, B., Davidian, M., & Tsiatis, A. A. 2020. *Dyntxregime: Methods for estimating optimal dynamic treatment regimes.* R package version 4.9.

Hu, Feifang, & Rosenberger, William F. 2006. *The theory of response-adaptive randomization in clinical trials.* Vol. 525. John Wiley & Sons.

Hu, Feifang, Zhang, Li-Xin, *et al.* 2004. Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *The annals of statistics*, **32**(1), 268–301.

Hu, Xinyu, Qian, Min, Cheng, Bin, & Cheung, Ying Kuen. 2021. Personalized policy learning using longitudinal mobile health data. *Journal of the american statistical association*, **116:533**, 410–420.

Hubbard, Alan E, Ahern, Jennifer, Fleischer, Nancy L, Van der Laan, Mark, Satariano, Sheri A, Jewell, Nicholas, Bruckner, Tim, & Satariano, William A. 2010. To gee or not to gee: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, 467–474.

Humphrey, Kyle. 2017. Using reinforcement learning to personalize dosing strategies in a simulated cancer trial with high dimensional data. *Master's thesis.*

Istepanian, Robert, Laxminarayan, Swamy, & Pattichis, Constantinos S. 2007. *M-health: Emerging mobile health systems.* Springer Science & Business Media.

Ivanova, Anastasia. 2003. A play-the-winner-type urn design with reduced variability. *Metrika*, **58**(1), 1–13.

Ivanova, Anastasia, & Durham, SD. 2000. Drop the loser rule. *University of north carolina at chapel hill technical report# tr-00-01.*

Ivanova, Anastasia, Rosenberger, William F, Durham, Stephen D, & Flournoy, Nancy. 2000. A birth and death urn for randomized clinical trials: asymptotic methods. *Sankhyā: the indian journal of statistics, series b*, 104–118.

Jaakkola, Tommi, Jordan, Michael I, & Singh, Satinder P. 1994. Convergence of stochastic iterative dynamic programming algorithms. *Pages 703–710 of: Advances in neural information processing systems.*

Jalalimanesh, Ammar, Haghighi, Hamidreza Shahabi, Ahmadi, Abbas, Hejazian, Hossein, & Soltani, Madjid. 2017a. Multi-objective optimization of radiotherapy: distributed q-learning and agent-based simulation. *Journal of experimental & theoretical artificial intelligence*, **29**(5), 1071–1086.

Jalalimanesh, Ammar, Haghighi, Hamidreza Shahabi, Ahmadi, Abbas, & Soltani, Madjid. 2017b. Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning. *Mathematics and computers in simulation*, **133**, 235–248.

Jamieson, Kevin, & Jain, Lalit. 2018. A bandit approach to multiple testing with false discovery control. *Page 3664–3674 of: 32nd Conference on Neural Information Processing Systems (NeurIPS).*

Jeffreys, H. 1961. Theory of probability (3rd edt.) oxford university press. *Mr0187257*, **432**.

Jennison, Christopher, & Turnbull, Bruce W. 2000. *Group sequential methods with applications to clinical trials.* Boca Raton, FL: Chapman & Hall/CRC Press.

Jennison, Christopher, & Turnbull, Bruce W. 2013. Interim monitoring of clinical trials: Decision theory, dynamic programming and optimal stopping. *Kuwait journal of science*, **40**(2).

Jiang, Fei, Jiang, Yong, Zhi, Hui, Dong, Yi, Li, Hao, Ma, Sufeng, Wang, Yilong, Dong, Qiang, Shen, Haipeng, & Wang, Yongjun. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, **2**(4), 230–243.

Johnson, Alistair EW, Ghassemi, Mohammad M, Nemati, Shamim, Niehaus, Katherine E, Clifton, David A, & Clifford, Gari D. 2016a. Machine learning and decision support in critical care. *Proceedings of the ieee*, **104**(2), 444–466.

Johnson, Alistair EW, Pollard, Tom J, Shen, Lu, Li-Wei, H Lehman, Feng, Mengling, Ghassemi, Mohammad, Moody, Benjamin, Szolovits, Peter, Celi, Leo Anthony, & Mark, Roger G. 2016b. Mimic-iii, a freely accessible critical care database. *Scientific data*, **3**(1), 1–9.

Jonsson, Anders. 2019. Deep reinforcement learning in medicine. *Kidney diseases*, **5**(1), 18–22.

Kaptein, Maurits. 2015. The use of thompson sampling to increase estimation precision. *Behavior research methods*, **47**(2), 409–423.

Kasari, Connie, Kaiser, Ann, Goods, Kelly, Nietfeld, Jennifer, Mathy, Pamela, Landa, Rebecca, Murphy, Susan, & Almirall, Daniel. 2014. Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the american academy of child & adolescent psychiatry*, **53**(6), 635–646.

Kass, Robert E, & Raftery, Adrian E. 1995. Bayes factors. *Journal of the american statistical association*, **90**(430), 773–795.

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

Kaufmann, Emilie, & Garivier, Aurélien. 2017. Learning the distribution with largest mean: two bandit frameworks. *Arxiv*, **abs/1702.00001**.

Keefe, Richard SE, Bilder, Robert M, Davis, Sonia M, Harvey, Philip D, Palmer, Barton W, Gold, James M, Meltzer, Herbert Y, Green, Michael F, Capuano, George, Stroup, T Scott, *et al.* 2007. Neurocognitive effects of antipsychotic medications in patients with chronic schizophrenia in the catie trial. *Archives of general psychiatry*, **64**(6), 633–647.

Kim, Gi-Soo, & Paik, Myunghee Cho. 2019. Contextual multi-armed bandit algorithm for semiparametric reward model. *In: Proceedings of the 36th international conference on machine learning, long beach, california, pmlr 97.*

Klasnja, Predrag, Hekler, Eric B, Shiffman, Saul, Boruvka, Audrey, Almirall, Daniel, Tewari, Ambuj, & Murphy, Susan A. 2015. Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. *Health psychology*, **34**(S), 1220.

Klasnja, Predrag, Smith, Shawna, Seewald, Nicholas J, Lee, Andy, Hall, Kelly, Luers, Brook, Hekler, Eric B, & Murphy, Susan A. 2019. Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of heartsteps. *Annals of behavioral medicine*, **53**(6), 573–582.

Kohavi, Ron, Deng, Alex, Frasca, Brian, Longbotham, Roger, Walker, Toby, & Xu, Ya. 2012. Trustworthy online controlled experiments: Five puzzling outcomes explained. *Pages 786–794 of: Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining.*

Kolesar, Peter. 1970. A markovian model for hospital admission scheduling. *Management science*, **16**(6), B–384.

Krause, Andreas, & Ong, Cheng S. 2011. Contextual gaussian process bandit optimization. *Pages 2447–2455 of: Advances in neural information processing systems*.

Kriemler, Susi, Meyer, Ursina, Martin, E, van Sluijs, Esther MF, Andersen, Lars Bo, & Martin, Brian W. 2011. Effect of school-based interventions on physical activity and fitness in children and adolescents: a review of reviews and systematic update. *British journal of sports medicine*, **45**(11), 923–930.

Krishnamurthy, Akshay, Wu, Zhiwei Steven, & Syrgkanis, Vasilis. 2018. Semiparametric contextual bandits. *arxiv preprint arxiv:1803.04204*.

Kroenke, Kurt, Spitzer, Robert L, & Williams, Janet BW. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, **16**(9), 606–613.

Kumar, Santosh, Nilsen, Wendy J, Abernethy, Amy, Atienza, Audie, Patrick, Kevin, Pavel, Misha, Riley, William T, Shar, Albert, Spring, Bonnie, Spruijt-Metz, Donna, *et al.* 2013. Mobile health technology evaluation: the mhealth evidence workshop. *American journal of preventive medicine*, **45**(2), 228–236.

Laber, Eric B, & Zhao, Ying-Qi. 2015. Tree-based methods for individualized treatment regimes. *Biometrika*, **102**(3), 501–514.

Laber, Eric B, Qian, Min, Lizotte, Dan J, Pelham, William E, & Murphy, Susan A. 2010. Statistical inference in dynamic treatment regimes. *arxiv preprint arxiv:1006.5831*.

Laber, Eric B, Lizotte, Daniel J, Qian, Min, Pelham, William E, & Murphy, Susan A. 2014a. Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, **8**(1), 1225.

Laber, Eric B, Linn, Kristin A, & Stefanski, Leonard A. 2014b. Interactive model building for q-learning. *Biometrika*, **101**(4), 831–847.

Laber, Eric B, Lizotte, Daniel J, & Ferguson, Bradley. 2014c. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, **70**(1), 53–61.

Lach, Denise. 2014. Challenges of interdisciplinary research: Reconciling qualitative and quantitative methods for understanding human–landscape systems. *Environmental management*, **53**(1), 88–93.

Lachin, John M. 1988. Statistical properties of randomization in clinical trials. *Controlled clinical trials*, **9**(4), 289–311.

Lachin, John M, Matts, John P, & Wei, LJ. 1988. Randomization in clinical trials: conclusions and recommendations. *Controlled clinical trials*, **9**(4), 365–374.

Lai, Tze Leung. 1987. Adaptive treatment allocation and the multi-armed bandit problem. *The annals of statistics*, 1091–1114.

Laird, Nan M, & Ware, James H. 1982. Random-effects models for longitudinal data. *Biometrics*, 963–974.

Lattimore, Tor, & Szepesvári, Csaba. 2020. *Bandit algorithms*. Cambridge University Press.

Lavori, Philip W, & Dawson, Ree. 2000. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the royal statistical society: Series a (statistics in society)*, **163**(1), 29–38.

Lavori, Philip W, & Dawson, Ree. 2004. Dynamic treatment regimes: practical design considerations. *Clinical trials*, **1**(1), 9–20.

LeCun, Y, *et al.* 1998. *Efficient backprop in: Neural networks: Tricks of the trade, this book is an outgrowth of a 1996 nips workshop*.

Lee, Hyun-Suk, Shen, Cong, Jordon, James, & van der Schaar, Mihaela. 2020. Contextual constrained learning for dose-finding clinical trials. *arxiv preprint arxiv:2001.02463*.

Lee, Juhee, Thall, Peter F, Ji, Yuan, & Müller, Peter. 2015a. Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *Journal of the american statistical association*, **110**(510), 711–722.

Lee, Min Kyung, Kim, Junsung, Forlizzi, Jodi, & Kiesler, Sara. 2015b. Personalization revisited: a reflective approach helps people better personalize health services and motivates them to increase physical activity. *Pages 743–754 of: Proceedings of the 2015 acm international joint conference on pervasive and ubiquitous computing*.

Lee, Paul H, Macfarlane, Duncan J, Lam, Tai Hing, & Stewart, Sunita M. 2011. Validity of the international physical activity questionnaire short form (ipaq-sf): A systematic review. *International journal of behavioral nutrition and physical activity*, **8**(1), 115.

Lei, Huitan, Nahum-Shani, Inbal, Lynch, Kevin, Oslin, David, & Murphy, Susan A. 2012. A" smart" design for building individualized treatment sequences. *Annual review of clinical psychology*, **8**, 21–48.

Lei, Huitian. 2016. *An online actor critic algorithm and a statistical decision procedure for personalizing intervention.* Ph.D. thesis, University of Michigan.

Lei, Huitian, Tewari, Ambuj, & Murphy, Susan A. 2017. An actor-critic contextual bandit algorithm for personalized mobile health interventions. *arxiv preprint arxiv:1706.09090*.

Levine, Nir, Crammer, Koby, & Mannor, Shie. 2017. Rotting bandits. *Pages 3074–3083 of: Advances in neural information processing systems*.

Li, Lihong. 2013. Generalized thompson sampling for contextual bandits. *arxiv preprint arxiv:1310.7163*.

Li, Lihong, Chu, Wei, Langford, John, & Schapire, Robert E. 2010. A contextual-bandit approach to personalized news article recommendation. *Pages 661–670 of: Proceedings of the 19th international conference on world wide web*.

Li, Lihong, Lu, Yu, & Zhou, Dengyong. 2017. Provably optimal algorithms for generalized linear contextual bandits. *arxiv preprint arxiv:1703.00048*.

Liang, Kung-Yee, & Zeger, Scott L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Liao, Peng, Klasnja, Predrag, Tewari, Ambuj, & Murphy, Susan A. 2015. Micro-randomized trials in mhealth. *arxiv preprint arxiv:1504.00238*.

Liao, Peng, Klasnja, Predrag, Tewari, Ambuj, & Murphy, Susan A. 2016. Sample size calculations for micro-randomized trials in mhealth. *Statistics in medicine*, **35**(12), 1944–1971.

Liao, Peng, Greenewald, Kristjan, Klasnja, Predrag, & Murphy, Susan. 2020. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the acm on interactive, mobile, wearable and ubiquitous technologies*, **4**(1), 1–22.

Libin, Pieter, Verstraeten, Timothy, Roijers, Diederik M, Wang, Wenjia, Theys, Kristof, & Nowe, Ann. 2019. Bayesian anytime m-top exploration. *Pages 1422–1428 of: 2019 ieee 31st international conference on tools with artificial intelligence (ictai)*. IEEE.

Libin, Pieter JK, Verstraeten, Timothy, Roijers, Diederik M, Grujic, Jelena, Theys, Kristof, Lemey, Philippe, & Nowé, Ann. 2018. Bayesian best-arm identification for selecting influenza mitigation strategies. *Pages 456–471 of: Joint european conference on machine learning and knowledge discovery in databases*. Springer.

Ling, Yuan, Hasan, Sadid A, Datla, Vivek, Qadir, Ashequl, Lee, Kathy, Liu, Joey, & Farri, Oladimeji. 2017. Diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning: A preliminary study. *Pages 271–285 of: Machine learning for healthcare conference*.

Linn, Kristin A, Laber, Eric B, & Stefanski, Leonard A. 2015. iqlearn: Interactive q-learning in r. *Journal of statistical software*, **64**(1).

Linn, Kristin A, Laber, Eric B, & Stefanski, Leonard A. 2017. Interactive q-learning for quantiles. *Journal of the american statistical association*, **112**(518), 638–649.

Liu, Xiaoxuan, Rivera, Samantha Cruz, Moher, David, Calvert, Melanie J, & Denniston, Alastair K. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *bmj*, **370**.

Liu, Ying, Logan, Brent, Liu, Ning, Xu, Zhiyuan, Tang, Jian, & Wang, Yangzhi. 2017. Deep reinforcement learning for dynamic treatment regimes on medical registry data. *Pages 380–385 of: 2017 ieee international conference on healthcare informatics (ichi)*. IEEE.

Liu, Ying, Wang, Yuanjia, Kosorok, Michael R, Zhao, Yingqi, & Zeng, Donglin. 2018. Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in medicine*, **37**(26), 3776–3788.

Liu, Yun-En, Mandel, Travis, Brunskill, Emma, & Popovic, Zoran. 2014. Trading off scientific knowledge and user learning with multi-armed bandits. *Pages 161–168 of: Edm*.

Lizotte, Daniel J, & Laber, Eric B. 2016. Multi-objective markov decision processes for data-driven decision support. *The journal of machine learning research*, **17**(1), 7378–7405.

Lizotte, Daniel J, Bowling, Michael, & Murphy, Susan A. 2012. Linear fitted-q iteration with multiple reward functions. *The journal of machine learning research*, **13**(1), 3253–3295.

Lomas, J Derek, Forlizzi, Jodi, Poonwala, Nikhil, Patel, Nirmal, Shodhan, Sharan, Patel, Kishan, Koedinger, Ken, & Brunskill, Emma. 2016. Interface design optimization as a multi-armed bandit problem. *Pages 4142–4153 of: Proceedings of the 2016 chi conference on human factors in computing systems*.

Luckett, Daniel J, Laber, Eric B, Kahkoska, Anna R, Maahs, David M, Mayer-Davis, Elizabeth, & Kosorok, Michael R. 2020. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the american statistical association*, **115**(530), 692–706.

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

Lyles, Courtney R, Handley, Margaret A, Ackerman, Sara L, Schillinger, Dean, Williams, Pamela, Westbrook, Marisa, Gourley, Gato, & Sarkar, Urmimala. 2019. Innovative implementation studies conducted in us safety net health care settings: a systematic review. *American journal of medical quality*, **34**(3), 293–306.

Maca, Jeff, Bhattacharya, Suman, Dragalin, Vladimir, Gallo, Paul, & Krams, Michael. 2006. Adaptive seamless phase ii/iii designs—background, operational aspects, and examples. *Drug information journal*, **40**(4), 463–473.

Maei, Hamid Reza, Szepesvári, Csaba, Bhatnagar, Shalabh, & Sutton, Richard S. 2010. Toward off-policy learning control with function approximation. *In: Icml*.

Mahajan, Rajiv., & Gupta, Kapil. 2010. Adaptive design clinical trials: Methodology, challenges and prospect. *Indian journal of pharmacology*, **42**(4), 201–207.

Malof, Jordan M, & Gaweda, Adam E. 2011. Optimizing drug therapy with reinforcement learning: The case of anemia management. *Pages 2088–2092 of: The 2011 international joint conference on neural networks*. IEEE.

Marquardt, Donald W, & Snee, Ronald D. 1975. Ridge regression in practice. *The american statistician*, **29**(1), 3–20.

Marsh, Lawrence C, & Cormier, David R. 2001. *Spline regression models*. Sage.

Martín-Guerrero, José D, Gomez, Faustino, Soria-Olivas, Emilio, Schmidhuber, Jürgen, Climente-Martí, Mónica, & Jiménez-Torres, N Víctor. 2009. A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert systems with applications*, **36**(6), 9737–9742.

McCarroll, Rebecca, Eyles, Helen, & Mhurchu, Cliona Ni. 2017. Effectiveness of mobile health (mhealth) interventions for promoting healthy eating in adults: A systematic review. *Preventive medicine*, **105**, 156–168.

McCarthy, Philip J. 1969. Pseudo-replication: Half samples. *Revue de l'institut international de statistique*, 239–264.

McCullagh, Peter. 2018. *Generalized linear models*. Routledge.

Means, Barbara, Toyama, Yuki, Murphy, Robert, Bakia, Marianne, & Jones, Karla. 2009. Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. *Centre for learning technology*.

Michie, Susan, Van Stralen, Maartje M, & West, Robert. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science*, **6**(1), 42.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, *et al.* 2015. Human-level control through deep reinforcement learning. *nature*, **518**(7540), 529–533.

Mohr, David C, Lyon, Aaron R, Lattie, Emily G, Reddy, Madhu, & Schueller, Stephen M. 2017. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. *Journal of medical internet research*, **19**(5), e153.

Moodie, Erica EM, & Richardson, Thomas S. 2010. Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian journal of statistics*, **37**(1), 126–146.

Moodie, Erica EM, Richardson, Thomas S, & Stephens, David A. 2007. Demystifying optimal dynamic treatment regimes. *Biometrics*, **63**(2), 447–455.

Moore, Brett L, Pyeatt, Larry D, Kulkarni, Vivekanand, Panousis, Periklis, Padrez, Kevin, & Doufas, Anthony G. 2014. Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers. *The journal of machine learning research*, **15**(1), 655–696.

Murphy, Susan A. 2003. Optimal dynamic treatment regimes. *Journal of the royal statistical society: Series b (statistical methodology)*, **65**(2), 331–355.

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

Murphy, Susan A. 2005b. A generalization error for q-learning. *Journal of machine learning research*, **6**(Jul), 1073–1097.

Murphy, Susan A, van der Laan, Mark J, Robins, James M, & Group, Conduct Problems Prevention Research. 2001. Marginal mean models for dynamic regimes. *Journal of the american statistical association*, **96**(456), 1410–1423.

Murphy, Susan A, Collins, Linda M, & Rush, A John. 2007. Customizing treatment to the patient: Adaptive treatment strategies. *Drug and alcohol dependence*, **88**(Suppl 2), S1.

Murphy, Susan A, Deng, Yanzhen, Laber, Eric B, Maei, Hamid Reza, Sutton, Richard S, & Witkiewitz, Katie. 2016. A batch, off-policy, actor-critic algorithm for optimizing the average reward. *arxiv preprint arxiv:1607.05047*.

Murray, Jennifer M, Brennan, Sarah F, French, David P, Patterson, Christopher C, Kee, Frank, & Hunter, Ruth F. 2017. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: a systematic review and meta-analysis. *Social science & medicine*, **192**, 125–133.

Murray, Thomas A, Yuan, Ying, & Thall, Peter F. 2018. A bayesian machine learning approach for optimizing dynamic treatment regimes. *Journal of the american statistical association*, **113**(523), 1255–1267.

Myung, Jay I, Cavagnaro, Daniel R, & Pitt, Mark A. 2013. A tutorial on adaptive design optimization. *Journal of mathematical psychology*, **57**(3-4), 53–67.

Nahum-Shani, Inbal, & Almirall, Daniel. 2019. An introduction to adaptive interventions and smart designs in education. ncser 2020-001. *National center for special education research*.

Nahum-Shani, Inbal, Hekler, Eric B, & Spruijt-Metz, Donna. 2015. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health psychology*, **34**(S), 1209.

Nahum-Shani, Inbal, Ertefaie, Ashkan, Lu, Xi, Lynch, Kevin G, McKay, James R, Oslin, David W, & Almirall, Daniel. 2017. A smart data analysis method for constructing adaptive treatment strategies for substance use disorders. *Addiction*, **112**(5), 901–909.

Nahum-Shani, Inbal, Smith, Shawna N, Spring, Bonnie J, Collins, Linda M, Witkiewitz, Katie, Tewari, Ambuj, & Murphy, Susan A. 2018. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of behavioral medicine*, **52**(6), 446–462.

Naughton, Felix. 2017. Delivering "just-in-time" smoking cessation support via mobile phones: current knowledge and future directions. *Nicotine & tobacco research*, **19**(3), 379–383.

Neugebauer, Romain, Fireman, Bruce, Roy, Jason A, O'Connor, Patrick J, & Selby, Joe V. 2012. Dynamic marginal structural modeling to evaluate the comparative effectiveness of more or less aggressive treatment intensification strategies in adults with type 2 diabetes. *Pharmacoepidemiology and drug safety*, **21**, 99–113.

Newton, Michael A, & Raftery, Adrian E. 1994. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the royal statistical society: Series b (methodological)*, **56**(1), 3–26.

Neyman, Jerzy S. 1923. On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of agricultural sciences*, **10**, 1–51.

Ng, Andrew Y, Russell, Stuart J, *et al.* 2000. Algorithms for inverse reinforcement learning. *Page 2 of: Icml*, vol. 1.

Ngo, Phuong D, Wei, Susan, Holubová, Anna, Muzik, Jan, & Godtliebsen, Fred. 2018. Reinforcement-learning optimal control for type-1 diabetes. *Pages 333–336 of: 2018 ieee embs international conference on biomedical & health informatics (bhi)*. IEEE.

Nie, Xinkun, Tian, Xiaoying, Taylor, Jonathan, & Zou, James. 2018. Why adaptively collected data have negative bias and how to correct for it. *Pages 1261–1269 of: International conference on artificial intelligence and statistics*. PMLR.

Nouri, Sarah S, Avila-Garcia, Patricia, Cemballi, Anupama Gunshekar, Sarkar, Urmimala, Aguilera, Adrian, & Lyles, Courtney Rees. 2019. Assessing mobile phone digital literacy and engagement in user-centered design in a diverse, safety-net population: mixed methods study. *Jmir mhealth and uhealth*, **7**(8), e14250.

Nyenhuis, Sharmilee, Ma, Jun, & Sharp, Lisa. 2017. Applying the com-b model to designing a tailored physical activity intervention for sedentary african american women with asthma. *Pages A3336–A3336 of: B38. asthma: A panoramic view*. American Thoracic Society.

Obermeyer, Ziad, & Lee, Thomas H. 2017. Lost in thought: the limits of the human mind and the future of medicine. *The new england journal of medicine*, **377**(13), 1209.

Ontanón, Santiago. 2013. The combinatorial multi-armed bandit problem and its application to real-time strategy games. *In: Ninth artificial intelligence and interactive digital entertainment conference*.

Ontanón, Santiago. 2017. Combinatorial multi-armed bandits for real-time strategy games. *Journal of artificial intelligence research*, **58**, 665–702.

O'Quigley, John, & Zohar, S. 2006. Experimental designs for phase i and phase i/ii dose-finding studies. *British journal of cancer*, **94**(5), 609–613.

O'Quigley, John, Pepe, Margaret, & Fisher, Lloyd. 1990. Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics*, **46**(1), 33–48.

Orellana, Liliana, Rotnitzky, Andrea, & Robins, James M. 2010. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. *The international journal of biostatistics*, **6**(2).

Organization, World Health, *et al.* 2005. *Preventing chronic diseases: a vital investment: Who global report*. World Health Organization.

Organization, World Health, *et al.* 2018. *Noncommunicable diseases country profiles 2018*. World Health Organization.

Owen, Art B, Eckles, Dean, *et al.* 2012. Bootstrapping data arrays of arbitrary order. *The annals of applied statistics*, **6**(3), 895–927.

Padmanabhan, Regina, Meskin, Nader, & Haddad, Wassim M. 2017. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Mathematical biosciences*, **293**, 11–20.

Pallmann, Philip, Bedding, Alun W, Choodari-Oskooei, Babak, Dimairo, Munyaradzi, Flight, Laura, Hampson, Lisa V, Holmes, Jane, Mander, Adrian P, Sydes, Matthew R, Villar, Sofía S, *et al.* 2018. Adaptive designs in clinical trials: why use them, and how to run and report them. *Bmc medicine*, **16**(1), 29.

Palumbo, Pasquale, Panunzi, Simona, & De Gaetano, Andrea. 2007. Qualitative behavior of a family of delay-differential models of the glucose-insulin system. *Discrete & continuous dynamical systems-b*, **7**(2), 399.

Pan, Wei. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**(1), 120–125.

Parbhoo, Sonali, Bogojeska, Jasmina, Zazzi, Maurizio, Roth, Volker, & Doshi-Velez, Finale. 2017. Combining kernel and model based learning for hiv therapy selection. *Amia summits on translational science proceedings*, **2017**, 239.

Paredes, Pablo, Gilad-Bachrach, Ran, Czerwinski, Mary, Roseway, Asta, Rowan, Kael, & Hernandez, Javier. 2014. Poptherapy: Coping with stress through pop-culture. *Pages 109–117 of: Proceedings of the 8th international conference on pervasive computing technologies for healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and ....

Park, Trevor, & Casella, George. 2008. The bayesian lasso. *Journal of the american statistical association*, **103**(482), 681–686.

Pascoe, Michaela, Bailey, Alan P, Craike, Melinda, Carter, Tim, Patten, Rhiannon, Stepto, Nigel, & Parker, Alexandra. 2020. Physical activity and exercise in youth mental health promotion: a scoping review. *Bmj open sport & exercise medicine*, **6**(1).

Patrick, Kevin, Raab, Fred, Adams, Marc, Dillon, Lindsay, Zabinski, Marion, Rock, Cheryl, Griswold, William, & Norman, Gregory. 2009. A text message-based intervention for weight loss: randomized controlled trial. *Journal of medical internet research*, **11**(1), e1.

Peizer, David B, & Pratt, John W. 1968. A normal approximation for binomial, f, beta, and other common, related tail probabilities, i. *Journal of the american statistical association*, **63**(324), 1416–1456.

Pelham, William E, Hoza, Betsy, Pillow, David R, Gnagy, Elizabeth M, Kipp, Heidi L, Greiner, Andrew R, Waschbusch, Daniel A, Trane, Sarah T, Greenhouse, Joel, Wolfson, Lara, *et al.* 2002. Effects of methyphenidate and expectancy on children with adhd: Behavior, academic performance, and attributions in a summer treatment program and regular classroom settings. *Journal of consulting and clinical psychology*, **70**(2), 320.

Pepe, Margaret Sullivan, & Anderson, Garnet L. 1994. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in statistics-simulation and computation*, **23**(4), 939–951.

Pfammatter, Angela Fidler, Nahum-Shani, Inbal, DeZelar, Margaret, Scanlan, Laura, McFadden, H Gene, Siddique, Juned, Hedeker, Donald, & Spring, Bonnie. 2019. Smart: study protocol for a sequential multiple assignment randomized controlled trial to optimize weight loss management. *Contemporary clinical trials*, **82**, 36–45.

Piette, John D, Farris, Karen B, Newman, Sean, An, Larry, Sussman, Jeremy, & Singh, Satinder. 2015. The potential impact of intelligent systems for mobile health self-management support: Monte carlo simulations of text message support for medication adherence. *Annals of behavioral medicine*, **49**(1), 84–94.

Pike-Burke, Ciara, & Grunewalder, Steffen. 2019. Recovering bandits. *Pages 14122–14131 of: Advances in neural information processing systems*.

Pineau, Joelle, Bellemare, Marc G, Rush, A John, Ghizaru, Adrian, & Murphy, Susan A. 2007. Constructing evidence-based treatment strategies using methods from computer science. *Drug and alcohol dependence*, **88**, S52–S60.

Prentice, Deborah A, & Miller, Dale T. 1992. When small effects are impressive. *Psychological bulletin*, **112**(1), 160.

Proschan, Michael A, Lan, KK Gordon, & Wittes, Janet Turk. 2006. *Statistical monitoring of clinical trials: a unified approach*. Springer Science & Business Media.

Puterman, Martin L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qian, Min, & Murphy, Susan A. 2011. Performance guarantees for individualized treatment rules. *Annals of statistics*, **39**(2), 1180.

Qian, Tianchen, Klasnja, Predrag, & Murphy, Susan A. 2020. Linear mixed models with endogenous covariates: modeling sequential treatment effects with application to a mobile health study. *Statistical science: a review journal of the institute of mathematical statistics*, **35**(3), 375.

Rabbi, Mashfiqui, Aung, Min Hane, Zhang, Mi, & Choudhury, Tanzeem. 2015. Mybehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. *Pages 707–718 of: Proceedings of the 2015 acm international joint conference on pervasive and ubiquitous computing*.

Rabbi, Mashfiqui, Klasnja, Predrag, Choudhury, Tanzeem, Tewari, Ambuj, & Murphy, Susan. 2019. Optimizing mhealth interventions with a bandit. *Pages 277–291 of: Digital phenotyping and mobile sensing*. Springer.

Rafferty, Anna, Ying, Huiji, & Williams, Joseph. 2019. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *Jedm— journal of educational data mining*, **11**(1), 47–79.

Raghu, Aniruddh, Komorowski, Matthieu, Celi, Leo Anthony, Szolovits, Peter, & Ghassemi, Marzyeh. 2017a. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. *arxiv preprint arxiv:1705.08422*.

Raghu, Aniruddh, Komorowski, Matthieu, Ahmed, Imran, Celi, Leo, Szolovits, Peter, & Ghassemi, Marzyeh. 2017b. Deep reinforcement learning for sepsis treatment. *arxiv preprint arxiv:1711.09602*.

Rajkomar, Alvin, Dean, Jeffrey, & Kohane, Isaac. 2019. Machine learning in medicine. *New england journal of medicine*, **380**(14), 1347–1358.

Rathbone, Amy Leigh, & Prescott, Julie. 2017. The use of mobile apps and sms messaging as physical and mental health interventions: systematic review. *Journal of medical internet research*, **19**(8), e295.

Raudenbush, Stephen W, & Bryk, Anthony S. 2002. *Hierarchical linear models: Applications and data analysis methods*. Vol. 1. sage.

Raynor, Hollie A, & Epstein, Leonard H. 2001. Dietary variety, energy regulation, and obesity. *Psychological bulletin*, **127**(3), 325.

Rehg, James M, Murphy, Susan A, & Kumar, Santosh. 2017. *Mobile health*. Springer.

Rigollet, Phillippe, & Hütter, Jan-Christian. 2015. High dimensional statistics. *Lecture notes for course 18s997*.

Riviere, Marie-Karelle, Yuan, Ying, Jourdan, Jacques-Henri, Dubois, Frédéric, & Zohar, Sarah. 2018. Phase i/ii dose-finding design for molecularly targeted agent: Plateau determination using adaptive randomization. *Statistical methods in medical research*, **27**(2), 466–479.

Robbins, Herbert. 1952. Some aspects of the sequential design of experiments. *Bulletin of the american mathematical society*, **58**(5), 527–535.

Roberts, Anna L, Fisher, Abigail, Smith, Lee, Heinrich, Malgorzata, & Potts, Henry WW. 2017. Digital health behaviour change interventions targeting physical activity and diet in cancer survivors: a systematic review and meta-analysis. *Journal of cancer survivorship*, **11**(6), 704–719.

Robins, James. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, **7**(9-12), 1393–1512.

Robins, James. 1992. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, **79**(2), 321–334.

Robins, James M. 1989. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on aids*, 113–159.

Robins, James M. 1994. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in statistics-theory and methods*, **23**(8), 2379–2412.

Robins, James M. 1997. Latent variable modeling and applications to causality. *Causal inference from complex longitudinal data*, 69–117.

Robins, James M. 2000. Marginal structural models versus structural nested models as tools for causal inference. *Pages 95–133 of: Statistical models in epidemiology, the environment, and clinical trials*. Springer.

Robins, James M. 2004. Optimal structural nested models for optimal sequential decisions. *Pages 189–326 of: Proceedings of the second seattle symposium in biostatistics*. Springer.

Robins, James M, Rotnitzky, Andrea, & Zhao, Lue Ping. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the american statistical association*, **89**(427), 846–866.

Romeo, Amelia, Edney, Sarah, Plotnikoff, Ronald, Curtis, Rachel, Ryan, Jillian, Sanders, Ilea, Crozier, Alyson, & Maher, Carol. 2019. Can smartphone apps increase physical activity? systematic review and meta-analysis. *Journal of medical internet research*, **21**(3), e12053.

Rose, Taylor, Barker, Mary, Jacob, Chandni Maria, Morrison, Leanne, Lawrence, Wendy, Strömmer, Sofia, Vogel, Christina, Woods-Townsend, Kathryn, Farrell, David, Inskip, Hazel, *et al.* 2017. A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. *Journal of adolescent health*, **61**(6), 669–677.

Rosenberger, William F, & Lachin, John M. 1993. The use of response-adaptive designs in clinical trials. *Controlled clinical trials*, **14**(6), 471–484.

Rosenberger, William F, & Lachin, John M. 2015. *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Rosenberger, William F, Stallard, Nigel, Ivanova, Anastasia, Harper, Cherice N, & Ricks, Michelle L. 2001. Optimal adaptive designs for binary response trials. *Biometrics*, **57**(3), 909–913.

Rosenblum, Michael, & Van Der Laan, Mark J. 2010. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The international journal of biostatistics*, **6**(2).

Rowland, Simon P, Fitzgerald, J Edward, Holme, Thomas, Powell, John, & McGregor, Alison. 2020. What is the clinical value of mhealth for patients? *Npj digital medicine*, **3**(1), 1–6.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, **66**(5), 688.

Rubin, Donald B. 1978. Bayesian inference for causal effects: The role of randomization. *The annals of statistics*, 34–58.

Rubin, Donald B. 1980. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the american statistical association*, **75**(371), 591–593.

Rubin, Donald B. 1981. The bayesian bootstrap. *The annals of statistics*, 130–134.

Rush, A John, Fava, Maurizio, Wisniewski, Stephen R, Lavori, Philip W, Trivedi, Madhukar H, Sackeim, Harold A, Thase, Michael E, Nierenberg, Andrew A, Quitkin, Frederic M, Kashner, T Michael, *et al.* 2004. Sequenced treatment alternatives to relieve depression (star* d): rationale and design. *Controlled clinical trials*, **25**(1), 119–142.

Russo, Daniel J., Roy, Benjamin Van, Kazerouni, Abbas, Osband, Ian, & Wen, Zheng. 2018. A tutorial on thompson sampling. *Foundations and trends® in machine learning*, **11**(1), 1–96.

Ruxton, Graeme D. 2006. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral ecology*, **17**(4), 688–690.

Santos, Maria Magdalena. 2013. Validation of the behavioral activation for depression scale-short form (bads-sf) with spanish-speaking latinos. *Theses and dissertations*.

Satlin, Andrew, Wang, Jinping, Logovinsky, Veronika, Berry, Scott, Swanson, Chad, Dhadda, Shobha, & Berry, Donald A. 2016. Design of a bayesian adaptive phase 2 proof-of-concept trial for ban2401, a putative disease-modifying monoclonal antibody for the treatment of alzheimer's disease. *Alzheimer's & dementia: Translational research & clinical interventions*, **2**(1), 1–12.

Schembre, Susan M, Liao, Yue, Robertson, Michael C, Dunton, Genevieve Fridlund, Kerr, Jacqueline, Haffey, Meghan E, Burnett, Taylor, Basen-Engquist, Karen, & Hicklen, Rachel S. 2018. Just-in-time feedback in diet and physical activity interventions: systematic review and practical design framework. *Journal of medical internet research*, **20**(3), e106.

Schueller, Stephen M, Hunter, John F, Figueroa, Caroline, & Aguilera, Adrian. 2019. Use of digital mental health for marginalized and underserved populations. *Current treatment options in psychiatry*, **6**(3), 243–255.

Schulte, Phillip J, Tsiatis, Anastasios A, Laber, Eric B, & Davidian, Marie. 2014. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the institute of mathematical statistics*, **29**(4), 640.

Segal, Avi, David, Yossi Ben, Williams, Joseph Jay, Gal, Kobi, & Shalom, Yaar. 2018. Combining difficulty ranking with multi-armed bandits to sequence educational content. *Pages 317–321 of: International conference on artificial intelligence in education*. Springer.

Shaikh, Hammad, Modiri, Arghavan, Williams, Joseph Jay, & Rafferty, Anna N. 2019. Balancing student success and inferring personalized effects in dynamic experiments. *In: Edm*.

Shen, Cong, Wang, Zhiyang, Villar, Sofia, & Van Der Schaar, Mihaela. 2020. Learning for dose allocation in adaptive clinical trials with safety constraints. *Pages 8730–8740 of: International conference on machine learning*. PMLR.

Shin, Jaehyeok, Ramdas, Aaditya, & Rinaldo, A. 2019. Are sample means in multi-armed bandits positively or negatively biased? *In: Neural Information Processing Systems (NeurIPS 2019)*.

Shin, Jaehyeok, Ramdas, Aaditya, & Rinaldo, A. 2020. On conditional versus marginal bias in multi-armed bandits. *In: Thirty-seventh International Conference on Machine Learning (ICML 2020)*.

Shortreed, Susan M, Laber, Eric, Lizotte, Daniel J, Stroup, T Scott, Pineau, Joelle, & Murphy, Susan A. 2011. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, **84**(1-2), 109–136.

Sibbald, Bonnie, & Roland, Martin. 1998. Understanding controlled trials. why are randomised controlled trials important? *Bmj: British medical journal*, **316**(7126), 201.

Silva, Anabela G, Simões, Patrícia, Queirós, Alexandra, Rodrigues, Mário, & Rocha, Nelson P. 2020. Mobile apps to quantify aspects of physical activity: a systematic review on its reliability and validity. *Journal of medical systems*, **44**(2), 51.

Silver, David, Schrittwieser, Julian, Simonyan, Karen, Antonoglou, Ioannis, Huang, Aja, Guez, Arthur, Hubert, Thomas, Baker, Lucas, Lai, Matthew, Bolton, Adrian, *et al.* 2017. Mastering the game of go without human knowledge. *nature*, **550**(7676), 354–359.

Smith, Adam L, & Villar, Sofía S. 2018. Bayesian adaptive bandit-based designs using the gittins index for multi-armed trials with normally distributed endpoints. *Journal of applied statistics*, **45**(6), 1052–1076.

Smucker, Byran, Krzywinski, Martin, & Altman, Naomi. 2018. Optimal experimental design. *Nature methods*, **15**(8), 559–560.

Song, Rui, Wang, Weiwei, Zeng, Donglin, & Kosorok, Michael R. 2015. Penalized q-learning for dynamic treatment regimens. *Statistica sinica*, **25**(3), 901.

Spangenberg, Lena, Braehler, Elmar, & Glaesmer, Heide. 2012. Identifying depression in the general population-a comparison of phq-9, phq-8 and phq-2. *Zeitschrift fur psychosomatische medizin und psychotherapie*, **58**(1), 3–10.

Spitzer, Robert L, Kroenke, Kurt, Williams, Janet BW, & Löwe, Bernd. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, **166**(10), 1092–1097.

Srinivas, Niranjan, Krause, Andreas, Kakade, Sham M, & Seeger, Matthias. 2009. Gaussian process optimization in the bandit setting: No regret and experimental design. *arxiv preprint arxiv:0912.3995*.

Srinivas, Niranjan, Krause, Andreas, Kakade, Sham M, & Seeger, Matthias W. 2012. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *Ieee transactions on information theory*, **58**(5), 3250–3265.

Stallard, Nigel, Hampson, Lisa, Benda, Norbert, Brannath, Werner, Burnett, Thomas, Friede, Tim, Kimani, Peter K, Koenig, Franz, Krisam, Johannes, Mozgunov, Pavel, *et al.* 2020. Efficient adaptive designs for clinical trials of interventions for covid-19. *Statistics in biopharmaceutical research*, **12**(4), 483–497.

Stuckey, Melanie I, Carter, Shawn W, & Knight, Emily. 2017. The role of smartphones in encouraging physical activity in adults. *International journal of general medicine*, **10**, 293.

Stupple, Aaron, Singerman, David, & Celi, Leo Anthony. 2019. The reproducibility crisis in the age of digital medicine. *Npj digital medicine*, **2**(1), 1–3.

Sugiyama, Masashi. 2015. *Statistical reinforcement learning: modern machine learning approaches*. CRC Press.

Sultan, Nabil. 2015. Reflective thoughts on the potential and challenges of wearable technology for healthcare provision and medical education. *International journal of information management*, **35**(5), 521–526.

Sun, Yilun, & Wang, Lu. 2020. Stochastic tree search for estimating optimal dynamic treatment regimes. *Journal of the american statistical association*, 1–12.

Sutton, Richard S, & Barto, Andrew G. 2018. *Reinforcement learning: An introduction*. MIT press.

Szepesvári, Csaba. 2010. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, **4**(1), 1–103.

Tao, Yebin, & Wang, Lu. 2017. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics*, **73**(1), 145–155.

Tao, Yebin, Wang, Lu, & Almirall, Daniel. 2018. Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *The annals of applied statistics*, **12**(3), 1914.

Taves, Donald R. 1974. Minimization: a new method of assigning patients to treatment and control groups. *Clinical pharmacology & therapeutics*, **15**(5), 443–453.

Tewari, Ambuj, & Murphy, Susan A. 2017. From ads to interventions: Contextual bandits in mobile health. *Pages 495–517 of: Mobile health*. Springer.

Thall, Peter F, & Cook, John D. 2004. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, **60**(3), 684—693.

Thall, Peter F, & Nguyen, Hoang Q. 2012. Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *Journal of biopharmaceutical statistics*, **22**(4), 785–801.

Thall, Peter F, & Wathen, J Kyle. 2007. Practical bayesian adaptive randomisation in clinical trials. *European journal of cancer*, **43**(5), 859–866.

Thall, Peter F, Millikan, Randall E, & Sung, Hsi-Guang. 2000. Evaluating multiple treatment courses in clinical trials. *Statistics in medicine*, **19**(8), 1011–1028.

Thall, Peter F, Sung, Hsi-Guang, & Estey, Elihu H. 2002. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the american statistical association*, **97**(457), 29–39.

Thall, Peter F, Wooten, Leiko H, Logothetis, Christopher J, Millikan, Randall E, & Tannir, Nizar M. 2007. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in medicine*, **26**(26), 4687–4702.

Thompson, William R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**(3/4), 285–294.

Thompson, William R. 1935. On the theory of apportionment. *American journal of mathematics*, **57**(2), 450–456.

Toga, Arthur W, Foster, Ian, Kesselman, Carl, Madduri, Ravi, Chard, Kyle, Deutsch, Eric W, Price, Nathan D, Glusman, Gustavo, Heavner, Benjamin D, Dinov, Ivo D, *et al.* 2015. Big biomedical data as the key resource for discovery science. *Journal of the american medical informatics association*, **22**(6), 1126–1131.

Toldov, Viktor, Clavier, Laurent, Loscrí, Valeria, & Mitton, Nathalie. 2016. A thompson sampling approach to channel exploration-exploitation problem in multihop cognitive radio networks. *Pages 1–6 of: 2016 ieee 27th annual international symposium on personal, indoor, and mobile radio communications (pimrc)*. IEEE.

Tomkins, Sabina, Liao, Peng, Yeung, Serena, Klasnja, Predrag, & Murphy, Susan. 2019. Intelligent pooling in thompson sampling for rapid personalization in mobile health. *In: Reinforcement learning for real life (rl4reallife) workshop in the 36th international conference on machine learning, longbeach, california, usa.*

Triantafyllidis, Andreas, Kondylakis, Haridimos, Votis, Konstantinos, Tzovaras, Dimitrios, Maglaveras, Nicos, & Rahimi, Kazem. 2019. Features, outcomes, and challenges in mobile health interventions for patients living with chronic diseases: A review of systematic reviews. *International journal of medical informatics*, **132**, 103984.

Triantafyllidis, Andreas K, & Tsanas, Athanasios. 2019. Applications of machine learning in real-life digital health interventions: review of the literature. *Journal of medical internet research*, **21**(4), e12286.

Tseng, Huan-Hsin, Luo, Yi, Cui, Sunan, Chien, Jen-Tzung, Ten Haken, Randall K, & El Naqa, Issam. 2017. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical physics*, **44**(12), 6690–6705.

Tsiatis, Anastasios A., Davidian, Marie D., Holloway, Shannon T., & Laber, Eric B. 2019. *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman & Hall/CRC.

Tsitsiklis, John N. 1994. Asynchronous stochastic approximation and q-learning. *Machine learning*, **16**(3), 185–202.

Urteaga, Iñigo, & Wiggins, Chris H. 2018. (sequential) importance sampling bandits. *arxiv preprint arxiv:1808.02933*.

Valko, Michal, Korda, Nathaniel, Munos, Rémi, Flaounas, Ilias, & Cristianini, Nelo. 2013. Finite-time analysis of kernelised contextual bandits. *arxiv preprint arxiv:1309.6869*.

Van Dantzig, Saskia, Geleijnse, Gijs, & Van Halteren, Aart Tijmen. 2013. Toward a persuasive mobile application to reduce sedentary behavior. *Personal and ubiquitous computing*, **17**(6), 1237–1246.

Van Otterlo, Martijn, & Wiering, Marco. 2012. Reinforcement learning and markov decision processes. *Pages 3–42 of: Reinforcement learning*. Springer.

Vansteelandt, Stijn, Joffe, Marshall, *et al.* 2014. Structural nested models and g-estimation: the partially realized promise. *Statistical science*, **29**(4), 707–731.

Vapnik, Vladimir, Golowich, Steven E, & Smola, Alex J. 1997. Support vector method for function approximation, regression estimation and signal processing. *Pages 281–287 of: Advances in neural information processing systems*.

Vellido, Alfredo, Ribas, Vicent, Morales, Carles, Sanmartín, Adolfo Ruiz, & Rodríguez, Juan Carlos Ruiz. 2018. Machine learning in critical care: state-of-the-art and a sepsis case study. *Biomedical engineering online*, **17**(1), 135.

Villar, Sofía S, Bowden, Jack, & Wason, James. 2015a. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the institute of mathematical statistics*, **30**(2), 199.

Villar, Sofía S, Wason, James, & Bowden, Jack. 2015b. Response-adaptive randomization for multi-arm clinical trials using the forward looking gittins index rule. *Biometrics*, **71**(4), 969–978.

Vincent, Robert. 2014. *Reinforcement learning in models of adaptive medical treatment strategies*. Ph.D. thesis, McGill University Libraries.

Voils, Corrine I, Chang, YunKyung, Crandell, Jamie, Leeman, Jennifer, Sandelowski, Margarete, & Maciejewski, Matthew L. 2012. Informing

the dosing of interventions in randomized trials. *Contemporary clinical trials*, **33**(6), 1225–1230.

Vollmer, Sebastian, Mateen, Bilal A, Bohner, Gergo, Király, Franz J, Ghani, Rayid, Jonsson, Pall, Cumbers, Sarah, Jonas, Adrian, McAllister, Katherine SL, Myles, Puja, *et al.* 2020. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj*, **368**.

Wages, Nolan A, Read, Paul W, & Petroni, Gina R. 2015. A phase i/ii adaptive design for heterogeneous groups with application to a stereotactic body radiation therapy trial. *Pharmaceutical statistics*, **14**(4), 302–310.

Wagner, Edward H, Austin, Brian T, Davis, Connie, Hindmarsh, Mike, Schaefer, Judith, & Bonomi, Amy. 2001. Improving chronic illness care: translating evidence into action. *Health affairs*, **20**(6), 64–78.

Wahed, Abdus S, & Tsiatis, Anastasios A. 2006. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika*, **93**(1), 163–177.

Wald, Abraham. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the american mathematical society*, **54**(3), 426–482.

Wallace, Michael, Moodie, Erica E M, Stephens, David A, Simoneau, Gabrielle, & Schulz, Juliana. 2020. *Dtrreg: Dtr estimation and inference via g-estimation, dynamic wols, q-learning, and dynamic weighted survival modeling (dwsurv)*. R package version 1.7.

Wallace, Michael P, & Moodie, Erica EM. 2015. Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, **71**(3), 636–644.

Walsh, Jane C, Corbett, Teresa, Hogan, Michael, Duggan, Jim, & McNamara, Abra. 2016. An mhealth intervention using a smartphone app to increase walking behavior in young adults: a pilot study. *Jmir mhealth and uhealth*, **4**(3), e109.

Walsh, Thomas J, Szita, István, Diuk, Carlos, & Littman, Michael L. 2012. Exploring compact reinforcement-learning representations with linear regression. *arxiv preprint arxiv:1205.2606*.

Wang, Lu, Rotnitzky, Andrea, Lin, Xihong, Millikan, Randall E, & Thall, Peter F. 2012. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the american statistical association*, **107**(498), 493–508.

Wang, Lu, Zhang, Wei, He, Xiaofeng, & Zha, Hongyuan. 2018. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. *Pages 2447–2456 of: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*.

Watkins, Christopher John Cornish Hellaby. 1989. *Learning from delayed rewards*. King's College, Cambridge.

Wei, Lee-Jen. 1978. The adaptive biased coin design for sequential experiments. *The annals of statistics*, 92–100.

Wei, LJ, & Durham, S. 1978. The randomized play-the-winner rule in medical trials. *Journal of the american statistical association*, **73**(364), 840–843.

Weisel, Kiona K, Fuhrmann, Lukas M, Berking, Matthias, Baumeister, Harald, Cuijpers, Pim, & Ebert, David D. 2019. Standalone smartphone apps for mental health—a systematic review and meta-analysis. *Npj digital medicine*, **2**(1), 1–10.

Welch, Bernard L. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, **34**(1/2), 28–35.

Wheeler, Graham M, Mander, Adrian P, Bedding, Alun, Brock, Kristian, Cornelius, Victoria, Grieve, Andrew P, Jaki, Thomas, Love, Sharon B, Weir, Christopher J, Yap, Christina, *et al.* 2019. How to design a dose-finding study using the continual reassessment method. *Bmc medical research methodology*, **19**(1), 1–15.

White, John. 2012. *Bandit algorithms for website optimization.* " O'Reilly Media, Inc.".

Whittle, Peter, *et al.* 1981. Arm-acquiring bandits. *The annals of probability*, **9**(2), 284–292.

WHO. 2018. Physical activity fact sheet. *Geneve: World health organization. retrieved february.* Accessed October, 2020.

Williams, Joseph Jay, Kim, Juho, Rafferty, Anna, Maldonado, Samuel, Gajos, Krzysztof Z, Lasecki, Walter S, & Heffernan, Neil. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. *Pages 379–388 of: Proceedings of the third (2016) acm conference on learning@ scale.*

Williams, Joseph Jay, Rafferty, Anna N, Ang, Andrew, Tingley, Dustin, Lasecki, Walter S, & Kim, Juho. 2017. Connecting instructors and learning scientists via collaborative dynamic experimentation. *Pages 3012–3018 of: Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems.*

Williams, Joseph Jay, Rafferty, Anna N, Tingley, Dustin, Ang, Andrew, Lasecki, Walter S, & Kim, Juho. 2018. Enhancing online problems through instructor-centered tools for randomized experiments. *Pages 1–12 of: Proceedings of the 2018 chi conference on human factors in computing systems.*

Williamson, S Faye, & Villar, Sofía S. 2020. A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*, **76**(1), 197–209.

Wilson, Timothy D, Centerbar, David B, Kermer, Deborah A, & Gilbert, Daniel T. 2005. The pleasures of uncertainty: prolonging positive moods in ways people do not anticipate. *Journal of personality and social psychology*, **88**(1), 5.

Xia, Yingce, Li, Haifang, Qin, Tao, Yu, Nenghai, & Liu, Tie-Yan. 2015. Thompson sampling for budgeted multi-armed bandits. *In: Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*.

Xiao, Yongling, Abrahamowicz, Michal, Moodie, Erica EM, Weber, Rainer, & Young, James. 2014. Flexible marginal structural models for estimating the cumulative effect of a time-dependent treatment on the hazard: reassessing the cardiovascular risks of didanosine treatment in the swiss hiv cohort study. *Journal of the american statistical association*, **109**(506), 455–464.

Xin, Jingyi, Chakraborty, Bibhas, & Laber, Eric B. 2012. *qlearn: Estimation and inference for q-learning*. R package version 1.0.

Xu, Jing, Yan, Xiaoxi, Figueroa, Caroline, Williams, Joseph Jay, & Chakraborty, Bibhas. 2020. Multi-level micro-randomized trial: Detecting the proximal effect of messages on physical activity. *arxiv preprint arxiv:2007.13741*.

Xu, Yanxun, Müller, Peter, Wahed, Abdus S, & Thall, Peter F. 2016. Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *Journal of the american statistical association*, **111**(515), 921–950.

Yang, Fanny, Ramdas, Aaditya, Jamieson, Kevin G, & Wainwright, Martin J. 2017. A framework for multi-a (rmed)/b (andit) testing with online fdr control. *Pages 5957–5966 of: Advances in neural information processing systems*.

Yao, Jiayu, Brunskill, Emma, Pan, Weiwei, Murphy, Susan, & Doshi-Velez, Finale. 2020. Power-constrained bandits. *arxiv preprint arxiv:2004.06230*.

Yasini, Sh, Naghibi-Sistani, MB, & Karimpour, A. 2009. Agent-based simulation for blood glucose control in diabetic patients. *International journal of applied science, engineering and technology*, **5**(1), 40–49.

Yauney, Gregory, & Shah, Pratik. 2018. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. *Pages 161–226 of: Machine learning for healthcare conference*.

Yi, Yanqing, & Wang, Xikui. 2011. Comparison of wald, score, and likelihood ratio tests for response adaptive designs. *Journal of statistical theory and applications*, **10**(4), 553–569.

Yin, Guosheng. 2012. *Clinical trial design: Bayesian and frequentist adaptive methods*. Vol. 876. John Wiley & Sons.

Yin, Guosheng, Li, Yisheng, & Ji, Yuan. 2006. Bayesian dose-finding in phase i/ii clinical trials using toxicity and efficacy odds ratios. *Biometrics*, **62**(3), 777–787.

Yom-Tov, Elad, Feraru, Guy, Kozdoba, Mark, Mannor, Shie, Tennenholtz, Moshe, & Hochberg, Irit. 2017. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical internet research*, **19**(10), e338.

Yu, Chao, Ren, Guoqi, & Liu, Jiming. 2019a. Deep inverse reinforcement learning for sepsis treatment. *Pages 1–3 of: 2019 ieee international conference on healthcare informatics (ichi)*. IEEE.

Yu, Chao, Liu, Jiming, & Nemati, Shamim. 2019b. Reinforcement learning in healthcare: A survey. *arxiv preprint arxiv:1908.08796*.

Yuan, Ying, Nguyen, Hoang Q., & Thall, Peter F. 2017. *Bayesian designs for phase i–ii clinical trials*. CRC Press.

Zajonc, Tristan. 2012. Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking. *Journal of the american statistical association*, **107**(497), 80–92.

Zame, William R, Bica, Ioana, Shen, Cong, Curth, Alicia, Lee, Hyun-Suk, Bailey, Stuart, Weatherall, James, Wright, David, Bretz, Frank, & van der Schaar, Mihaela. 2020. Machine learning for clinical trials in the era of covid-19. *Statistics in biopharmaceutical research*, **12**(4), 506–517.

Zang, Yong, Lee, J Jack, & Yuan, Ying. 2014. Adaptive designs for identifying optimal biological dose for molecularly targeted agents. *Clinical trials*, **11**(3), 319–327. PMID: 24844841.

Zelen, M. 1974. The randomization and stratification of patients to clinical trials. *Journal of chronic diseases*, **27**(7), 365 – 375.

Zelen, Marvin. 1969. Play the winner rule and the controlled clinical trial. *Journal of the american statistical association*, **64**(325), 131–146.

Zhang, Baqun, & Zhang, Min. 2018. C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics*, **74**(3), 891–899.

Zhang, Baqun, Tsiatis, Anastasios A, Davidian, Marie, Zhang, Min, & Laber, Eric. 2012a. Estimating optimal treatment regimes from a classification perspective. *Stat*, **1**(1), 103–114.

Zhang, Baqun, Tsiatis, Anastasios A, Laber, Eric B, & Davidian, Marie. 2012b. A robust method for estimating optimal treatment regimes. *Biometrics*, **68**(4), 1010–1018.

Zhang, Baqun, Tsiatis, Anastasios A, Laber, Eric B, & Davidian, Marie. 2013. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, **100**(3), 681–694.

Zhang, Chong, Chen, Jingxiang, Fu, Haoda, He, Xuanyao, Zhao, Ying-Qi, & Liu, Yufeng. 2020a. Multicategory outcome weighted margin-based learning for estimating individualized treatment rules. *Statistica sinica*.

Zhang, Kelly W., Janson, Lucas, & Murphy, Susan A. 2020b. Inference for batched bandits. *In: Neural Information Processing Systems (NeurIPS 2020)*.

Zhang, Kelly W, Janson, Lucas, & Murphy, Susan A. 2020c. Inference for batched bandits. *arxiv preprint arxiv:2002.03217*.

Zhang, Li-Xin, Chan, Wai Sum, Cheung, Siu Hung, & Hu, Feifang. 2007. A generalized drop-the-loser urn for clinical trials with delayed responses. *Statistica sinica*, **17**(1), 387–409.

Zhang, Wei, Sargent, Daniel J., & Mandrekar, Sumithra. 2006. An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in medicine*, **25**(14), 2365–2383.

Zhang, Zhongheng. 2016. Multiple imputation with multivariate imputation by chained equation (mice) package. *Annals of translational medicine*, **4**(2).

Zhao, Ying-Qi, Zeng, Donglin, Laber, Eric B, & Kosorok, Michael R. 2015. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the american statistical association*, **110**(510), 583–598.

Zhao, Yingqi, Zeng, Donglin, Rush, A John, & Kosorok, Michael R. 2012. Estimating individualized treatment rules using outcome weighted learning. *Journal of the american statistical association*, **107**(499), 1106–1118.

Zhao, Yufan, Kosorok, Michael R, & Zeng, Donglin. 2009. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, **28**(26), 3294–3315.

Zhao, Yufan, Zeng, Donglin, Socinski, Mark A, & Kosorok, Michael R. 2011. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, **67**(4), 1422–1433.

Zhou, Dongruo, Li, Lihong, & Gu, Quanquan. 2019. Neural contextual bandits with upper confidence bound-based exploration. *arxiv preprint arxiv:1911.04462*.

Zhou, Mo, Mintz, Yonatan, Fukuoka, Yoshimi, Goldberg, Ken, Flowers, Elena, Kaminsky, Philip, Castillejo, Alejandro, & Aswani, Anil. 2018. Personalizing mobile fitness apps using reinforcement learning. *In: Ceur workshop proceedings*, vol. 2068. NIH Public Access.

Zhou, Xin, Mayer-Hamblett, Nicole, Khan, Umer, & Kosorok, Michael R. 2017. Residual weighted learning for estimating individualized treatment rules. *Journal of the american statistical association*, **112**(517), 169–187.

Zhu, Feiyun, Guo, Jun, Li, Ruoyu, & Huang, Junzhou. 2018. Robust actor-critic contextual bandit for mobile health (mhealth) interventions. *Proceedings of the 2018 acm international conference on bioinformatics, computational biology, and health informatics*.