Evaluating the Use of Generalized Dynamic Weighted Ordinary Least Squares for Individualized HIV Treatment Strategies

Larry Dong

ANNEXE – PAGE D'EN-TETE et 4ème de couverture

2nd year of the Master Sciences, Technologies, Santé :
mention Santé Publique - Parcours Public Health Data Science

Year 2019-2020

Evaluating the Use of Generalized Dynamic Weighted Ordinary Least Squares for Individualized HIV Treatment Strategies

From January 6, 2020 to August 15, 2020

McGill University
Department of Epidemiology, Biostatistics and Occupational Health

Purvis Hall
1020 Pine Avenue West
Montreal, QC
H3G 1A2

Master of Science in Public Health Data Science
Master of Science in Biostatistics

Discussed on the Wednesday, June 3, 2020
By Larry Dong

EUR DPH Graduate program funding supported by PIA3

Example of back cover

Abstract

Dynamic treatment regimes (DTR) are a statistical paradigm in personalized medicine which aims to optimize the outcome of long-term treatments. At its simplest, a DTR can optimize for a decision rule which consists of a single treatment; such DTRs are called individualized treatment rules (ITR) and they are often used in optimizing short-term rather than long-term rewards. DTR estimation methods such as dynamic weighted ordinary least squares (dWOLS) offer desirable theoretical advantages such as double robustness of parameter estimates in the decision rules. A recent extension of dWOLS called generalized dWOLS can now accommodate categorical treatments in the estimation of optimal treatment strategies. An application of this novel method will be demonstrated on HIV-infected individuals called low immunological responders (LIR), who are characterized by their low CD4+ T cell counts despite receiving antiretroviral treatment. The administration of Interleukin 7 (IL-7) injections has been shown to increase the concentration of CD4 cells in LIRs, but the optimal number of injections has yet to be determined. In this project, an ITR will be devised to maximize the duration where the CD4+ load is above a healthy threshold (500 cells/µL) while preventing the administration of unnecessary injections.

Keywords:

Individualized treatment rule, HIV, Interleukin 7, personalized medicine

Address:
Université de Bordeaux
Institut de Santé Publique d'Epidémiologie et de Développement
146 rue Léo Saignat
CS 61292
33076 Bordeaux cedex
www.u-bordeaux.fr

Contents

List of Figures

List of Tables

Research Environment

This research project was conducted jointly at the Department of Epidemiology, Biostatistics and Occupational Health at McGill University and ISPED, School of Public Health at the University of Bordeaux in conjunction with Bordeaux Population Health (BPH). This work was done under the supervision of Dr. Erica Moodie from McGill University and Dr. Rodolphe Thiébaut from the University of Bordeaux. Also affiliated with the Centre de Recherche Mathematiques (CRM), Dr. Moodie has numerous active collaborations with researchers at McGill and in other Canadian and international institutions; Dr. Thiébaut is affiliated with Inserm and Inria, two well-known research institutions in France, also leads the Statistics in Systems Biology and Translational Medicine (SISTM) team at the BPH. The BPH is a research center based in Bordeaux which consists of 11 research teams, all of which are working within the context of public health; the SISTM team focuses on the development of statistical methods for the analysis of clinical and biological data. A more detailed organization chart alongside its administrative structure and staff is provided in the following link: `https://www.bordeaux-population-health.center/wp-content/uploads/2018/11/Organigramme_U1219_31-08-2018_complet-sign%C3%A9.pdf`.

This completion of this project was made possible thanks to the Masters Excellence scholarship from the Institute of Data Valorisation (IVADO), Frontenac scholarship from the Fonds de recherche nature et technologies (FRQNT) and Graduate Mobility Award from the McGill Faculty of Medicine. Lastly, the assistance and knowledge of Dr. Laura Villain, postdoctoral researcher at SISTM and ISPED, was very beneficial in understanding and familiarization of the research topic.

1 Introduction

In personalized medicine or precision medicine, the central paradigm lies in adopting *patient-centric* medical practices rather than a *disease-centric* approach [37]. Due to the heterogeneity of diseased individuals, the effects of different available treatments can largely vary from patient to another. Although these *rule-of-thumb* approaches are often simpler and easier to implement in practice, different methods are continuously being developed to assist medical decision makers to tailor treatment plans according to subject-level information. For instance, particularly in the clinical management of chronic illnesses, many things to consider on top of patient health include compliance to medical care, ease of financial burden caused by treatments and reduction in adverse effects amongst many others [5, 37]. The evolving nature of ailments such as cancer, depression and substance abuse amongst many other long-term conditions often calls for treatments to be adapted to patient response and overall well-being.

While public health officials have raised awareness to the benefits of population-based health interventions from societal, economical and political perspectives, the potential benefits of personalized medicine has garnered interest in many areas of research over the past few years [12, 37]. For medical decision makers, interest lies in determining the optimal treatments for individual patients in the attempt of providing more adequate, personal and overall better health care for them. For statisticians, they are interested in better understanding and quantifying the risks and benefits of different treatment plans through the development robust yet easily interpretable data-driven methods. That being said, it is also important for both statisticians and decision makers to be able to quantify the uncertainty of such estimation methods in a theoretical framework and to assess their reliability when the proposed model is wrong. From a quantitative perspective, the challenges of this emerging medical framework lie in constructing such a rigorous yet readily applicable framework using data from clinical trials or observational studies.

1.1 Objectives

In this project, a method within the dynamic treatment regime (DTR) framework called generalized dynamic ordinary least squares (G-dWOLS) will be applied in a population of people living with HIV (PLWH) to investigate the benefits administration of exogenous Interleukin-7 (IL-7) injections. HIV is characterized by a depletion of CD4+ T cells (CD4), which are responsible for the proper functioning of the immune system [8]. In the absence of treatment, this loss in CD4 cells can lead to acquired immunodeficiency syndrome (AIDS), which makes those who have been infected vulnerable to opportunistic infections such as pneumonia, malaria and bacterial infections amongst many other ailments [28]. Highly active antiretroviral therapy (HAART), the most common treatment for HIV, inhibits the replication of the virus and it is often followed by a proliferation of CD4 T cells [8]. It has been shown that failure to reconstitute of CD4 cells in HIV-infected individuals to a threshold greater or equal to 500 cells$/\mu L$ is associated with a higher mortality rate and increased risk in developing opportunistic infections [24, 26, 33]. However, despite having no detectable viral load, 15% to 30% of individuals receiving HAART are known as poor or low immunological responders (LIR)

due to their inability to increase their CD4 level [26, 16].

The goal of this analysis is to devise a short-term or static treatment rule that optimizes the number of injections while preventing the administration of unnecessary injections using G-dWOLS [36]. There are costs – both financial and clinical (such as treatment fatigue and risk of side effects) – to injections, and so it is of interest to find the smallest number of injections needed to ensure CD4 cell counts lie above 500 cells/μL. While the DTR framework is able to optimize long-term outcomes that require multiple treatments, estimation of individualized treatment rules (ITR) is simply a DTR where the decision vector consists of a single treatment and it can be used in situations where it is assumed that there are no delayed effects of the treatment(s). In this analysis, the focus of the optimization problem will be on short term outcomes primarily because the benefits of IL-7 injections on the increase in CD4 count appear to be immediate.

1.2 Interleukin-7 and the INSPIRE Studies

Interleukin 7 (IL-7) or Recombinant Human Interleukin 7 (r-hIL-7) is a cytokine that plays an essential role in the survival and maintenance of CD4 cells [22, 45, 46, 50]. While IL-7 is naturally produced in stromal and epithelial cells of the bone marrow and thymus, multiple studies, both in humans and in animals, suggest that the administration of IL-7 leads to the reconstitution of CD4 cell counts [46, 27]. Amongst these studies, clinical studies INSPIRE 2 and 3 have been conducted where researchers examined the benefits of administering repeated injections of IL-7 in LIRs [51]. According to biological activity and tolerance towards injections, Lévy et al. have also shown that the ideal dosage for IL-7 injections was 20μg/kg, and hence this dose was used in clinical trials such as INSPIRE 2 and 3 [23].

INSPIRE 2 was a single-arm clinical trial where participants were drawn from an adult population of PLWH who have been receiving HAART for over a year and exhibiting suboptimal CD4 counts, i.e. between 100 and 400 cells/μL [18, 23]. The protocol entails providing a cycle of 3 injections of IL-7 at one week intervals, monitoring patient T cell response quarterly and readministering another cycle after 12 months of follow-up to maintain the CD4 load above 500 cells/μL [23, 51]. INSPIRE 3 was a clinical study where participants were randomized into a CYT107 arm and a control arm at a 3:1 ratio. With eligibility criteria defined similarly to those used in INSPIRE 2, the clinical protocol in the CYT107[1] entailed beginning patients with a cycle of 3 injections, and repeating a cycle of 3 injections if patients presented a CD4 cell count < 550 cells/μL at any quarterly evaluations. Patients in the control arm had their CD4 load measured for one year without any IL-7 injections. After one year, similarly to participants in the CYT107 arm, a cycle of injections was administered if patients in the control arm presented a CD4 cell count below 550 cells/μL [51].

The main finding of the INSPIRE 2 and 3 clinical trials is that participants exhibited T cell proliferation after receiving injections of the cytokine [51]. Recent studies have attempted to model CD4 trajectories and optimize IL-7 treatment using the data from the INSPIRE trials [50, 53, 34, 18]. Thiébaut et al. have investigated the restoration of CD4 cells attributable to exogenous IL-7 injections by mod-

[1]CYT107 refers to encoding of r-hIL-7 by Cytheris, a biopharmaceutical company providing the injection contents [6, 51].

elling CD4 dynamics using mechanistic models [18, 50]. Villain et al. have focused on devising injection protocols by predicting future instances where patient CD4 load may fall under 500 cells/μL [53]. Pazin et al. have used optimal control theory to determine an optimal number of IL-7 injections that would allow LIRs to maintain a healthy CD4 load [34].

1.3 Research Question

Because the INSPIRE 2 and 3 protocol calls for injections if the CD4 load falls below 550 cells/μL at quarterly evaluations, the optimal number of injections will be determined at 90 day time intervals [50]. In other words, the results from the statistical analysis using ITR modelling will provide the ideal number of injections according to a patient's profile and response to previous injections. The research question of interest is to determine the optimal number of injections while preventing the administration of unnecessary injections. To accommodate the trade-off between immune response utility and number of injection administered, respectively represented by variable U^g and U^i whose definitions are detailed in section 2.6.2, the outcome variable U^η is guided by a hyperparameter η chosen between 0 and 1 inclusively. For instance, if $\eta = 0$, the outcome will be the negative number of injections administered and the optimization problem would be to minimize the provision of treatment regardless of a patient's immune response. Likewise, if $\eta = 1$, all the weight would be shifted to U^g and the goal would now be to maximize the duration where CD4 is above the 500 cells/μL threshold.

The idea behind the definition outcome variable $U^\eta(\cdot)$ shares similar goal with the Q-TWiST method, short for Quality-adjusted Time Without Symptoms of disease or Toxicity of treatment. The Q-TWiST method leverages the trade-off between quality and quantity of life, especially in the clinical management of chronic illnesses and palliative care; recent work has investigated and compared different treatment plans in metastatic prostate cancer, breast cancer and childhood malignancies amongst many other ailments [14, 15, 41]. The overarching goal behind Q-TWiST is to define a single outcome which captures utilities from different sources. The Q-TWiST method, first introduced by Gelber et al., defines a single outcome of interest using a weighted sum of three utilities: treatment toxicity, time spent devoid of symptoms or side effects and time after relapse [14]. For instance, patients who are more prone to adverse effects may put a stronger emphasis on symptom-free time whereas others may prefer stronger treatments if it decreases the chances of illness relapse. The Q-TWiST method provides a framework to compare different treatment options for patients with different clinical preferences.

1.4 Research Hypothesis

My hypothesis regarding this research project is that the dWOLS analysis would provide valuable insight on the INSPIRE data. However, the clinical implications from the results of this statistical analysis can be limited due to low sample size in certain subpopulations. By construction of the outcome variable, it should be expected that, for $\eta = 0$, most, if not all, participants are recommended no injections regardless of other treatment-specific information. However, if more weight is attributed to a larger value of η, it is possible that not all patient-stages are recommended the administration

of a cycle of 3 injections; it could be interesting to investigate the reason behind a recommendation of fewer than 3 injections in such participants. Given that the adaptation of the INSPIRE data for an ITR analysis is adequate, a potentially interesting finding of this project would be the variation of treatment recommendation between patients with different characteristics.

2 Methods

2.1 Overview

Dynamic treatment regimes (DTR) are a statistical paradigm in personalized medicine that optimizes an outcome of interest through sequential decision making. Formally speaking, a DTR is a function that receives patient history as input and outputs an optimal decision vector which itself is comprised of one or multiple decisions, depending on the nature of the problem. At its simplest, a DTR is an estimation procedure that optimizes for a single decision; as mentioned earlier, such single-stage DTRs are referred as individual treatment rules (ITRs) which will be discussed more extensively in subsection 2.3. For instance, an example of a single-stage clinical problem could be deciding whether to discharge or not patients in the emergency room depending on their ailments and the severity of their symptoms.

In general, the procedure of optimizing for multi-stage decision problem requires knowledge of the underlying data-generating mechanism. However, although they are often unknown, they can be estimated using data-driven methods. This is where evidence-based medicine meets statistics, where the development of more reliable theoretical framework is fundamental in finding optimal treatment regimes and especially in quantifying their uncertainty. The primary goals of DTRs are twofold: comparing expected utilities of different deterministic treatments and obtaining optimal personalized treatment plans for patients [5]. One of the biggest strengths of the DTR framework is its ability to optimize long-term outcomes which involve medical interventions that need to be performed in a cascading fashion. In a multi-stage setting, the set of decision rules can be perceived as a medical protocol which can be adapted by monitoring patient well-being as they progress through a specific treatment strategy. Recent work has extensively studied the properties of statistical procedures such as Q-Learning and G-estimation [29, 30, 31, 32, 39]. In this work, we will be focusing on dynamic weighted ordinary least squares due to its recent extension to accommodate categorical treatments and on ITRs, since they are better suited for the problem that we wish to study.

2.2 Notation

When talking about ITRs, some terms and notation are worth introducing to clarify important concepts. The notation presented below assumes a single-stage setting because the focus will be on ITRs. The overarching ideas behind multi-stage DTRs are detailed in section 2.5 titled *Longitudinal Framework*.

Definition 2.1 (Outcome)**.** An **outcome** Y is a measure of a patient's response towards a particular treatment or sequence of treatments. In particular, it attempts to quantify a patient's overall wellness and it should also be sensitive enough to capture changes in a patient's state of well-being. Without loss of generality, it is defined to be non-negative and, as such, a larger value of Y represents a better overall state.

Definition 2.2 (Treatment)**.** A **treatment** $A \in \mathcal{A}$ is a medical intervention whose goal is to improve a patient's quality of life and well-being. Although it is commonplace in DTR literature to assume

A to be a binary variable for simplicity purposes, i.e. $\mathcal{A} = \{0, 1\}$, recent progress in DTR estimation methods allows A_j to be categorical or continuous[2][38, 42].

When there are m treatment options to be selected from where $m > 2$, this works assumes without loss of generality that $\mathcal{A} = \{a_\ell\}_{\ell=1}^{m}$.

Definition 2.3 (Covariates)**.** The **covariates** of patient i are represented by the p-dimensional row vector $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ and refer to each subject's non-treatment information that can influence the treatment effect on the measured outcome[3]. An $(n \times p)$ matrix $\boldsymbol{X}$ can be used to represent the collection of n different p-dimensional individual-specific covariate information.

$$\boldsymbol{X} = \begin{bmatrix} — & \boldsymbol{X}_1 & — \\ & \vdots & \\ — & \boldsymbol{X}_n & — \end{bmatrix} = \begin{bmatrix} X_{11} & \ldots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \ldots & X_{np} \end{bmatrix}$$

Definition 2.4 (Propensity score)**.** The **propensity score** of an individual is denoted by $\pi(\boldsymbol{x}, a) = P(A = a \,|\, \boldsymbol{X} = \boldsymbol{x})$ and represents the probability of receiving a particular treatment $a \in \mathcal{A}$ [40].

It is worth noting that, in statistical literature, the term propensity score was initially coined for binary treatments or exposures. The term generalized propensity score is more commonly used in situations where treatments can be categorical or continuous [1, 40].

When working with DTRs, positing simpler models such as linear regressions to estimate contrasts of utilities offers simpler interpretability of estimated parameters. While this is advantageous for medical decision makers from a clinical perspective, it also facilitates the understanding of statistical properties of estimators and provides a foundation for the expansion of theoretical work. The *blip* and *regret* functions defined below will be useful when discussing about DTRs or ITRs, alongside *treatment-free* and *treatment* models when talking about linear expressions in the statistical modelling procedure of DTRs.

Definition 2.5 (Blip function)**.** A *blip* function denoted by $\gamma(\boldsymbol{x}, a; \boldsymbol{\psi})$ is defined to be the expected gain in outcome if treatment $a \in \mathcal{A}$ were to be chosen instead of a reference treatment a^{ref}.

$$\gamma(\boldsymbol{x}, a; \boldsymbol{\psi}) = \mathbb{E}\left[Y\left(\boldsymbol{x}, a\right) - Y\left(\boldsymbol{x}, a^{\text{ref}}\right) \,|\, \boldsymbol{X} = \boldsymbol{x}, A = a\right]$$

In other words, the blip function characterizes the expected gain in the outcome variable upon providing a different available intervention strategy.

Definition 2.6 (Regret function)**.** The *regret* function $\mu(\boldsymbol{x}, a)$ is defined to be the expected loss from receiving some treatment $a \in \mathcal{A} = \{0, 1\}$ instead of the optimal treatment a^{opt}.

$$\begin{aligned} \mu(\boldsymbol{x}, a) &= \mathbb{E}\left[Y\left(\boldsymbol{x}, a^{\text{opt}}\right) - Y\left(\boldsymbol{x}, a\right) \,|\, \boldsymbol{X} = \boldsymbol{x}, A = a\right] \\ &= \gamma(\boldsymbol{x}, a^{\text{opt}}; \boldsymbol{\psi}) - \gamma_j(\boldsymbol{x}, a; \boldsymbol{\psi}) \end{aligned}$$

[2]Some literature in DTRs uses $A_j \in \{-1, 1\}$ to denote a binary treatment. In this work, the notation $A \in \{0, 1\}$ will be used as it is more intuitive to use $A = 0$ to represent an absence of treatment.

[3]In this work, any vector will be represented by a bold symbol.

In fact, when solving for optimal ITRs, the regret function is not needed in the estimation procedure. However, when handling longitudinal data, the sequential process of finding optimal regimes requires the regret function.

Definition 2.7 (Treatment model)**.** The treatment model or propensity score model is the model that predicts the probability that a patient with covariates $\boldsymbol{x}$ receives some treatment $a \in \mathcal{A}$. The treatment model denoted by $\mathbb{E}[A \mid \boldsymbol{X}^{\alpha}; \alpha]$ is often estimated by a (multinomial) logistic regression model.

Definition 2.8 (Treatment-free model)**.** When modelling for the outcome variable Y, the model is often written as the sum of two expressions: the treatment-free model and the blip function. In other words, the treatment-free model is the portion of the outcome model that is independent of the treatment.

$$\underbrace{\mathbb{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}, A = a; \boldsymbol{\beta}, \boldsymbol{\psi}\right]}_{\text{outcome model}} = \underbrace{G(\boldsymbol{x}^{\beta}; \boldsymbol{\beta})}_{\text{treatment-free model}} + \underbrace{\gamma(\boldsymbol{x}^{\psi}, a; \boldsymbol{\psi})}_{\text{blip function}}$$

In summary, the nomenclature for the models in the outcome model is due to the clear separation of linear expressions due to the terms which interact with the treatment variable of interest. Observe that the three models detailed above – treatment, treatment-free and blip model – are parameterized by coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$. While more detail will be provided regarding estimation methods for estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$, regression-based approaches for DTR estimation posits linear models for the latter ones, i.e. $G(\boldsymbol{x}^{\beta}; \boldsymbol{\beta}) = \boldsymbol{x}^{\beta}\boldsymbol{\beta}$ and $\gamma(\boldsymbol{x}^{\psi}, a; \boldsymbol{\psi}) = \boldsymbol{x}^{\psi}\boldsymbol{\psi}$. Also note that the superscripts β and ψ are to label explanatory variables with respect to their respective "submodel" within the outcome model; $\boldsymbol{X}^{\beta}$ and $\boldsymbol{X}^{\psi}$ are typically subsets of covariates $\boldsymbol{X}$.

2.3 Individualized Treatment Rules for Myopic Regimes

Especially in the context of chronic illnesses, patients require ongoing and long-term medical attention to treat their ailments. As mentioned earlier, one of the main advantages of the DTR framework is its ability to handle longitudinal data and solve for globally optimal decision rules [19]. However, ITRs can be used in the optimization of long-term outcomes through the maximization of short-term or immediate rewards. The overarching idea behind this approach bears strong similarities with the greedy algorithm whereby locally optimal decisions are made in the attempt of maximizing an overall or terminal outcome [19]. The essence of this method stems from the trade-off between the simpler formulation of the problem-solving framework and a potentially non-optimal solution that is relatively close to the best one [19]. While such global optimal treatment strategies are preferable to statically optimal regimes, the benefits in solving for myopic treatment regimes are its simpler interpretation and a more lightweight statistical estimation procedure. The process of estimating optimal ITRs, also known as statically optimal rules, can involve less model-based extrapolation while yielding more precise (i.e. narrower) confidence intervals [36].

In DTRs, rather than evaluating regimes through a sequence of decision rules, one of many ways to approach the study design to treat every stage as a single observation. In doing so, the sequential nature of interventions is eliminated and the optimization procedure ensues with the maximization

of short-term rewards. Just like in the greedy algorithm, the main disadvantage in employing this simpler study design is such an optimization procedure may not optimize an outcome in the long run. However, unlike in applications of the greedy algorithm, the primary benefit of employing such a simpler study design is not the gain in computational efficacy but rather easier interpretation of solutions [19, 36]. This is particularly useful in the design of myopic treatment regimes, which are multi-stage problems characterized by an assumption of no delayed treatment effects. In other words, from an algorithmic perspective, such problems have the ability to output globally optimal decision rules formed via a series of locally optimized actions, since individual decisions have no long-term consequences.

2.3.1 Induced Correlation and Variance-Covariance Structures

When solving multi-stage decision problems using a myopic regime study design, each participant contributes n_i observations into an agglomerated dataset; in a clinical context, each observation can be referred to as patient-stage. Because this data rearrangement procedure creates many observations for a single individual, an intuitive assumption is to suppose that there is some degree of correlation between measurements contributed by a same person. Many statistical methods have been developed to accurately estimate parameters while acknowledging underlying correlation structures. A general framework provided in Diggle et al. allows statistical inference to be performed on regression coefficients under the assumption of different correlation structures [7].

In the presence of correlation within observations from a same individual (or group or cluster), while each participants' outcomes can still be assumed to be independent of one and another, it may be naive to assume that the outcome of their own treatment stages are mutually independent. To account for the correlation of repeated measurements, different working variance-covariance structures can be assumed for the conditional variance of $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, denoted by $V_i = \text{Var}(Y_i)$[4]. In fact, one particular correlation structure called the independence correlation structure assumes that there is no underlying relationship between data points from a same group or source. When evaluating the use of myopic regimes, this can be naive and unrealistic to assume that the outcome of their own treatment stages are mutually independent. However, individuals each go through n_i treatment stages and independence between subjects can still be assumed, independence between outcomes from different subjects can still be assumed, but perhaps not between observations from a same individual.

Formally put, individuals identified by a subscript $1 \leq i \leq n$, are associated to a set of tuples or sequence of observations denoted by $\{(Y_{ij}, A_{ij}, \boldsymbol{X}_{ij})\}_{j=1}^{n_i}$. Since the general linear framework would be relevant for regression-based approaches in DTR estimation, the following general linear model can be postulated for the outcome model.

$$Y_{ij} = \boldsymbol{x}_{ij}^{\beta}\boldsymbol{\beta} + a_{ij}\boldsymbol{x}_{ij}^{\psi}\boldsymbol{\psi} + \epsilon_{ij}$$

[4] In the analysis of longitudinal data, V_i is often referred as the working variance matrix.

It is important to highlight the assumption of inter-subject independence, showcased in the following expression[5].

$$\boldsymbol{\epsilon_i} \perp \boldsymbol{\epsilon_{i'}} \qquad \text{for } 1 \leq i, i' \leq n, \quad i \neq i'$$

In standard linear regression, ϵ_{ij} would be independent and identically distributed for all $1 \leq i \leq n, 1 \leq j \leq m$. As such, the general linear model for handle sequential observations from a same subject. Using the matrix notation $(\boldsymbol{Y}_i, \boldsymbol{A}_i, \boldsymbol{X}_i)$ to denote the random vector or matrices for subject i measurements, $\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i, \boldsymbol{A}_i$ follows a multivariate normal distribution.

$$\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i, \boldsymbol{A}_i \sim \mathcal{N}\left(\boldsymbol{X}^{\beta}\boldsymbol{\beta} + \boldsymbol{A}_i\boldsymbol{X}^{\psi}\boldsymbol{\psi},\, \sigma^2 V\right)$$

For instance, a simple yet naive assumption would be to assume complete independence between all observations regardless of the potential underlying correlation. The variance-covariance matrix V would bear the following form.

$$V_i = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Many other covariance structures are possible to adjust for the correlation of repeated measurements. An exchangeable structure, also known as uniform correlation model, proposes that $\text{cor}\,(Y_{ij}, Y_{ij'}) = \rho$ for $1 \leq j, j' \leq n_i$ such that $j \neq j'$. In other words, any pair of non-identical measurement from same subject will be correlated in the same way. Under exchangeability, we have the following V_i structure.

$$V_i = \begin{bmatrix} \sigma^2 & \rho & \dots & \rho \\ \rho & \sigma^2 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \dots & \dots & \sigma^2 \end{bmatrix}$$

Another commonly used structure for the variance-covariance matrix is the exponential correlation model, also known as first-order autoregressive model. Provided that measurements $Y_{i1}, \dots, Y_{in_i}$ where taken at timestamps $t_1, \dots, t_{n_i}$, this type correlation relationship has that observations sampled more closely together are more related.

[5] $\boldsymbol{\epsilon_i} = \left(\epsilon_{i1}, \dots, \epsilon_{in_i}\right)^{\top}$

$$V_i = \begin{bmatrix} \sigma^2 & \rho & \rho^2 & \rho^3 & \dots & \rho^{n_i-1} \\ \rho & \sigma^2 & \rho & \rho^2 & \vdots & \vdots \\ \rho^2 & \rho & \sigma^2 & \rho & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \rho^{n_i-4} & \dots & \sigma^2 \end{bmatrix}$$

2.3.2 Empirical Estimation of Standard Errors

The main challenge of specifying a covariance structure is that it is often difficult to uncover the true form of the variance-covariance matrix V_i. However, for a linear model $\mathbb{E}[Y \mid X] = X\boldsymbol{\beta}$, the standard errors of $\widehat{\boldsymbol{\beta}}$ can be estimated empirically as follows [7].

$$\widehat{\text{Var}\left(\widehat{\beta}\right)} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i^\top V_i^{-1} \boldsymbol{x}_i\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i^\top V_i^{-1} \text{Var}(Y_i) V_i^{-1} \boldsymbol{x}_i\right) \left(\sum_{i=1}^{n} \boldsymbol{x}_i^\top V_i^{-1} \boldsymbol{x}_i\right)^{-1}$$

This formula is commonly referred as the robust variance estimator or sandwich estimator due to the presence of identical term visually surrounding another expression. Observe that, if the true form of the $\text{Var}(Y_i)$ is in fact the posited V_i, we have that $\text{Var}\left(\widehat{\beta}\right) = \left(\sum_{i=1}^{n} \boldsymbol{x}_i^\top V_i^{-1} \boldsymbol{x}_i\right)^{-1}$. However, if the assumed variance-covariance structure is incorrect, we can still evaluate the expression above by plugging in the emperically estimated V_i and $\widehat{\text{Var}(Y_i)}$, which can be estimated via the squared residuals. By doing so, the empirally estimated variance of the coefficients of interests are still consistent, i.e. $\widehat{\text{Var}(\widehat{\beta})} \to \text{Var}(\widehat{\beta})$ [17]. However, while the robustness of standard errors in the face of the misspecification of V_i permits the "safe" use of the independence correlation structure, this comes as a cost of reduced efficiency [7]. In cases where there is underlying correlation between observations, one can simply continue with linear regression and accommodate for correlation in the estimation of standard errors by using the sandwich formula above.

However, the procedure for estimating the standard errors of blip coefficients for an ITR needs to incorporate the error in estimating generalized propensity scores. In other words, while the sandwich formula has been shown to be consistent for V_i^*, the dWOLS analysis incorporates its own weights in the estimation procedure of DTRs. These weights are themselves estimated values, which means that an adjustment to the robust standard error formula is required for them to provide valid inference.

2.3.3 Bootstrap Estimation of Standard Errors and Confidence Intervals

Bootstrapping is a statistical method first introduced by Bradley Efron in 1979 which aims to estimate parameters of interest through iterative resampling of existing data [9, 52]. A highly computational and automatic method, this empirical estimation method can be used to estimate parameters, standard errors and confidence interval bounds and its theoretical; its theoretical simplicity and ease in implementation are desirable. Its wide range of applications stems from the lack of asserting any

distributional assumption on the data at hand when using the bootstrap procedure.

For a vector-valued random variable $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ from an underlying distribution F, given observed values $\boldsymbol{z} = (z_1, \ldots, z_n)$, a bootstrap sample indicated by $z^* = (z_1^*, \ldots, z_n^*)$ is obtained by sampling from the realizations $\{z_i\}_{i=1}^n$ with replacement. Because sampling is done with replacement, values z_i can appear anywhere from 0 to n times in the bootstrap sample z^*. Specifically in the context of DTRs, Chakraborty et al. have investigated the necessary changes in the bootstrap algorithm to be able to perform valid inference on quantities [4].

Using the notation employed in Efron and Tibshirani (1993) and the introductory notions presented in the same work, the bootstrap estimation method for a parameter of interest θ begins with a known estimator for θ, $\widehat{\theta} = s(\boldsymbol{x})$ where $s(\cdot)$ is some plug-in function of the observations $\boldsymbol{x}$ [52]. Theoretical properties of $\widehat{\theta}$ such as its expectation and variance can be evaluated using the underlying distribution F, but also estimated using bootstrap methods.

$$\mathbb{E}(\widehat{\theta}) = \int s(\boldsymbol{x}) \mathrm{d}F \qquad \text{Var}\left(\widehat{\theta}\right) = \int \left\{ s(\boldsymbol{x}) - \mathbb{E}(\widehat{\theta}) \right\} \mathrm{d}F$$

With bootstrapping, an empirical distribution $\widehat{F}$ is used instead of the true probability distribution F. It follows that the bootstrap replication of $\widehat{\theta}$ is obtained by evaluating $s(\cdot)$ on the resampled version of $\boldsymbol{x}^*$.

$$\widehat{\theta}^* = s(\boldsymbol{x}^*)$$

The standard error of $\widehat{\theta}$, denoted by $se(\widehat{\theta})$, can be estimated by performing the resampling procedure multiple times and assessing the spread of all obtained values of θ^*. For a total of B iterations, using $\widehat{\theta}_b^*$ to represent the bth bootstrap replication of $\widehat{\theta}$, the bootstrap estimate of the standard error of $\widehat{\theta}$, indicated by $\widehat{se}_B$ can be obtained as followed.

$$\widehat{se}_B = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\theta}_b^* - \overline{\theta}^* \right)^2} \qquad \text{where } \overline{\theta^*} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\theta}_b^*$$

Most importantly, it has been shown that $\lim_{B \to \infty} \widehat{se}_B = se(\widehat{\theta})$. Likewise, estimates of probability quantiles can also greatly benefit from bootstrap methods. Since the estimator $\widehat{\theta}$ is itself a random variable, say that is has some distribution $F_{\widehat{\theta}}$. Consider the $100 \cdot \alpha$th quantile of $F_{\widehat{\theta}}$ denoted by $\widehat{\theta}^{*(\alpha)}$; an appropriate estimate for this quantity would be simply the $100 \cdot \alpha$th ordered value out of a set of B estimates $\{\widehat{\theta}_b^*\}_{b=1}^B$, all of which are bootstrap replication of $\widehat{\theta}$. This proves to be useful in estimating $100(1-\alpha)\%$ confidence interval, which can be done empirically as followed.

$$\left[\widehat{\theta}^{(\alpha)}, \widehat{\theta}^{(1-\alpha)} \right] \approx \left[\widehat{\theta}_B^{*(\alpha)}, \widehat{\theta}_B^{*(1-\alpha)} \right]$$

For instance, say that 1000 iterations of bootstrap resampling are performed. Then, the estimated value for the 20th percentile of $F_{\widehat{\theta}}$ is the 200th largest value of the 1000 bootstrap replications. More sophisticated quantile estimation methods using bootstrap resampling are detailed in Chapters 13

and 14 [52]. More accurate intervals can be achieved at the cost of positing assumptions on the behaviour of the estimator $\widehat{\theta}$, but it is common to use the nonparametric estimator in 2.3.3 due to its ease in execution. Within the context of linear model, especially in DTRs, it is necessary to perform inference on the obtained blip coefficient estimates.

Standard error of linear coefficients are typically obtained using regression-based formulas, ignoring potential correlation between observations. Since the bootstrap estimation process is generalizable for any data-generating mechanism F, in which correlation may occur, the estimate values obtained using bootstrap resampling are in fact consistent. On top of being easily implementable, this empirical estimation method offers theoretical benefits when estimating parameters of interest.

2.4 Regression-Based Approaches for DTR Estimation

When working with DTRs, positing simpler models to model contrasts of utilities offers simpler interpretability of estimated parameters [5]. Recent work has examined the use of more flexible and sophisticated models such as decision trees [48] and deep neural network [25]. In this section, the focus will be on three regression-based approaches: Q-Learning, G-estimation, dynamic weighted ordinary least squares (dWOLS) and a generalized version of dWOLS.

2.4.1 Q-Learning and G-Estimation

Primary difference between Q-Learning and G-estimation is the compromise between methodological complexity and theoretical guarantees. Proper inference on blip coefficients in Q-Learning requires a correct outcome model specification whereas the blip parameters obtained from G-estimation boasts the doubly-robust property. In single-stage Q-Learning, the estimation process is nothing more than solving a linear regression.

$$\mathbb{E}\left[Y \mid \boldsymbol{X}, A\right] = \boldsymbol{X}^{\beta}\boldsymbol{\beta} + A\boldsymbol{X}^{\psi}\boldsymbol{\psi}$$

Although this method may seem overly simplistic in a single-stage decision problem, the main advantages of Q-Learning is its recursive formulation to optimize for sequential decision making. When the functional form is not correctly specified, desirable statistical properties (e.g. unbiasedness, low variance, consistency, etc.) are not guaranteed; this property is known as single-robust. When comparing treatments strategies, the treatment-free model parameters are considered "nuisance" since interest lies in quantifying the relative quality of treatment regimes to each other. G-estimation is a doubly-robust method that puts more attention on the blip coefficients. Rather than requiring the entire outcome model to be correctly specified, consistency of estimators of blip model parameters can still be achieved in light of misspecification of the treatment-free model. The double robustness of $\widehat{\boldsymbol{\psi}}$ obtained through G-estimation allows consistency to be achieved using a correctly specific treatment model. Positing a correct treatment-free model calls for knowledge regarding the data-generating mechanism for the outcome model, which is, in most cases, impossible to obtain. However, in many observational settings, the allocation of treatment options can be controlled by researchers, hence allowing correct specification of the treatment model much more feasible. For

single-stage decision problems, the G-estimation procedure can be summarized in the following steps.

1. Select (possibly identical) subsets of explanatory variables $\boldsymbol{X}^{\alpha}, \boldsymbol{X}^{\beta}$ and $\boldsymbol{X}^{\psi}$ from $\boldsymbol{X}$ for the treatment model, the treatment-free model and the blip model respectively.
2. Estimate the probability of receiving treatment $\pi(A, \boldsymbol{x}) = \mathbb{E}\left[A \,|\, \boldsymbol{X}; \boldsymbol{\alpha}\right]$.
3. Define $G(\boldsymbol{x}; \boldsymbol{\beta}) = Y - \gamma(\boldsymbol{x}, a_j; \boldsymbol{\psi}) = Y - a\boldsymbol{x}^{\psi}\boldsymbol{\psi}$ and $S(A, \boldsymbol{X}) = \frac{\partial}{\partial \boldsymbol{\psi}}\gamma(\boldsymbol{x}, A; \boldsymbol{\psi}) = A\boldsymbol{x}^{\psi}$.
4. Posit a model for $\mathbb{E}\left[G(\boldsymbol{X}; \boldsymbol{\psi}) \,|\, \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\beta}\right]$; in most cases, a linear model is be used.

$$\mathbb{E}\left[G(\boldsymbol{X}; \boldsymbol{\psi}) \,|\, \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\beta}\right] = \boldsymbol{x}^{\beta}\boldsymbol{\beta}$$

5. Obtain the estimates $\widehat{\boldsymbol{\psi}}$ by solving the following score function $U(\boldsymbol{\psi}) = \mathbf{0}$ defined below.

$$\begin{aligned} U(\boldsymbol{\psi}) &= \sum_{i=1}^{n} \left(G(\boldsymbol{\psi}) - \mathbb{E}\left[G(\boldsymbol{x}; \boldsymbol{\beta}\right]\right) \left(S(A; \boldsymbol{\alpha}) - \mathbb{E}\left[S(A, \boldsymbol{X}) \,|\, \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\alpha}\right]\right) \\ &= \sum_{i=1}^{n} \left(Y_i - \boldsymbol{x}_i^{\beta}\boldsymbol{\beta} - a_i\boldsymbol{x}_i^{\psi}\boldsymbol{\psi}\right) \left(\{a_i - \pi(a_i, \boldsymbol{x}; \boldsymbol{\alpha})\}\boldsymbol{x}^{\psi}\right) \end{aligned}$$

In step 2, the most common method in estimating the propensity score is via logistic regression. Also it is worth highlighting that a closed-form solutions for $\boldsymbol{\psi}$ can be derived by working with the score function in step 5 since $U(\boldsymbol{\psi})$ is linear in $\boldsymbol{\psi}$. The most important result of G-estimation is the double robustness of $\widehat{\boldsymbol{\psi}}$. However, although G-estimation offers theoretical advantages, its challenging presentation and intimidating implementation may deter researchers from using this estimation method. A similar and equally statistically robust method to G-estimation is dynamic weighted ordinary least squares, which also guarantees consistency of blip estimators $\widehat{psibf}$ under correct specification of either the treatment-free or treatment model.

2.4.2 Dynamical Weighted Ordinary Least Squares

Dynamic weighted ordinary least squares (dWOLS) is a weighted regression-based approach for the estimation of optimal treatment regimes first introduced by Wallace et Moodie (2015) [54]. This subsection highlights the theoretical underpinnings and practical considerations detailed in the article. The main advantages in using dWOLS compared to other methods are threefold: its relatively intuitive computational implementation, its statistically robust estimation procedure and its ability to accomodate categorical treatments. Similarly to G-estimation, dWOLS provides doubly-robust blip parameters by utilizing treatment-specific information. This appealing property relies on having weights $w(\boldsymbol{x}, a)$ associated to each individual-observations satisfying the balancing property below. On top of boasting the same statistical robustness as G-estimation, the dWOLS algorithm has the potential of being less daunting in implementation due to the familiar function form of score function, as it is equivalent to solving a weighted least squares. This equivalence in methodological procedure allows a relatively straightforward implementation the dWOLS algorithm using built-in regression functions with weights adhering to the balancing property.

Theorem 2.1 (Balancing Property)**.** Given that weights satisfying following equation

$$\pi\left(\boldsymbol{x}, 1\right) w\left(1, \boldsymbol{x}\right) = \pi(\boldsymbol{x}, 0) w\left(0, \boldsymbol{x}\right)$$

the dWOLS estimation procedures ensures that the blip parameter estimates $\boldsymbol{\psi}$ are consistent where $\pi\left(\boldsymbol{x}, a\right) = P\left(A = a \,|\, \boldsymbol{X} = \boldsymbol{x}\right)$.

Many weights satisfy the balancing property above. A common choice of weights is the inverse probability treatment weights (IPTW) where $w(a, \boldsymbol{x}) = \{\pi(\boldsymbol{x}, a)\}^{-1}$. Other family of weights that satisfy the balancing property are detailed in Wallace et Moodie (2015) [54]. In practice, when using dWOLS in estimation DTRs, the selection of the weights comes down to researcher's preference. However, the standard errors estimates can vary depending the form of weights since the asymptotic variance of blip parameters depend on $w(\cdot)$ [42, 54]. In summary, the dWOLS estimation method for ITRs is illustrated in the 5 following steps.

1. Select (possibly identical) subsets of covariates $\boldsymbol{X}^{\alpha}, \boldsymbol{X}^{\beta}$ and $\boldsymbol{X}^{\psi}$ from $\boldsymbol{X}$ for the treatment model, the treatment-free model and the blip function respectively.
2. Propose a treatment model $\mathbb{E}\left[A \,|\, \boldsymbol{X}\right]$ and define a weight w such that the balancing condition is satisfied.
3. Posit treatment-free model $f(\cdot)$ in $\mathbb{E}\left[Y \,|\, \boldsymbol{X} = \boldsymbol{x}, A = a\right] = f\left(\boldsymbol{X}^{\beta}\boldsymbol{\beta}\right) + A\boldsymbol{X}^{\psi}\boldsymbol{\psi}$. Typically, linear regressions are used due to their appealing statistical properties and simplicity, i.e. $f(\boldsymbol{X}^{\beta}, \boldsymbol{\beta}) = \boldsymbol{X}^{\beta}\boldsymbol{\beta}$.
4. Solve the following system of estimating equations to obtain the blip function parameter estimates $\widehat{\boldsymbol{\psi}}$:

$$\mathbf{0}_{(r+q)\times n} = \sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{X}_i^{\beta\top} \\ A_i \boldsymbol{X}_i^{\psi\top} \end{pmatrix} w_i \left(Y_i - \boldsymbol{X}_i^{\beta}\boldsymbol{\beta} - A_i \boldsymbol{X}_i^{\psi}\boldsymbol{\psi}\right)$$

   Note that a closed-formed expression for $\widehat{\boldsymbol{\psi}}$ is available.
5. Evaluate the optimal treatment plan for each subject given estimates $\widehat{\boldsymbol{\psi}}$ as followed.

$$A^{\text{opt}} = \begin{cases} 1 & \text{if } \boldsymbol{x}^{\psi}\widehat{\boldsymbol{\psi}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Recent work has extended in the dWOLS literature to handle survival outcomes and continuous treatments [42, 43]. This project investigates the application of generalized dWOLS which can accommodate multinomial or categorical treatment. While the estimation procedure for single-stage categorical decision problems closely follows the outline of the dWOLS procedure above, the adaptations to several components warrant a more in-depth inspection into the additional theoretical considerations and results of using the generalized version of dWOLS.

2.4.3 Generalized dWOLS for Categorical Treatments

The description of the dWOLS estimation procedure above assumes a binary treatment variable A for proof of concept purposes. However, a recent extension of dWOLS, also known as generalized dWOLS (G-dWOLS), is able to accommodate for categorical treatments or interventions with more than 2 treatment options possible. This is especially relevant for the data at hand provided from the INSPIRE studies, which will be addressed in section 1.2. Given a set of possible treatment options $\mathcal{A} = \{a_1, \ldots, a_m\}$ for some $m > 2$, the estimators produced by dWOLS are still doubly-robust under similar conditions as in the binary case. The two main adaptations in the dWOLS structure to accomodate categorical treatment is its balancing property and outcome model.

Balancing property – As exhibited in the balancing property theorem above, doubly-robust blip parameters stem from providing appropriate weights to each observation in the weighted least squares algorithm. The formulation of the theorem is general, in that $\pi(\boldsymbol{x}, a)\, w(\boldsymbol{x}, a)$ needs to bear the same value for all interventions $a \in \mathcal{A}$. Examples of multinomial weights adhering to the balancing property that will be abroaded in this work include inverse probability of treatment weights (IPTW) and overlap weights (W-O):

- Inverse probability of treatment weights (IPTW)

$$w(\boldsymbol{x}, a) = \frac{1}{\pi(\boldsymbol{x}, a)}$$

- Overlap weights (W-O):

$$w(\boldsymbol{x}, a) = \frac{1/\pi(\boldsymbol{x}, a)}{\sum_{\ell=1}^{m} 1/\pi(\boldsymbol{x}, a_\ell)}$$

  Notice that the overlap weights are in fact IPTW divided by a stabilizing term $\sum_{\ell=1}^{m} \frac{1}{\pi(\boldsymbol{x}, a_\ell)}$ that is solely depends on $\boldsymbol{x}$ and not any treatment options.

Careful attention must be made when positing structure on weights $w(\cdot, a)$; direct extension of weights in binary context to the categorical setting may not necessarily adhere to the balancing property [42]. In Schultz and Moodie (2020), other weighting families for categorical treatments adhering to the balancing property are also available, but will not be addressed in this paper [42].

Outcome model and score function – On top of the adaptations of weights, careful attention must be given to the outcome model and score function. In the binary case, the blip function can be understood as an expected gain in utility or reward when receiving treatment $A = 1$ compared to $A = 0$; it is insinuated that $A = 0$ is the baseline or reference treatment. However, when more than 2 treatments are possible, adaptations to blip function and hence the score function must be made. Keeping the treatment-free model untouched, the sum of $m - 1$ contrast terms will compose the "new" blip function. In other words, a linear term $\boldsymbol{x}^{\psi}\psi_\ell$ must be posited for each non-reference treatment ℓ. Simply put, when A is a categorical variable where $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$, assume without loss of generality that a_1 is chosen to be the reference treatment. The blip function can be written as a sum of treatment contrasts, each of which represent an expected gain in utility compared to the

counterfactual situation where a_1 were the assignment treatment.

$$\gamma(\boldsymbol{x}^{\psi}, a) = \sum_{\ell=2}^{m} \mathbb{1}_{a=a_\ell}\, \boldsymbol{x}^{\psi}\boldsymbol{\psi}_\ell$$

Notice that $\mathbb{1}_{a=a_\ell} = 1$ for at most one $\ell \in \{2, \ldots, m\}$. Closed-form solutions for the blip function parameters $\boldsymbol{\psi}$ can be obtained by solving a vector-valued score function similarly to the one provided in section 2.4.2.

$$\mathbf{0}_{p\times 1} = \sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{X}_i^{\beta} \\ \mathbb{1}_{a=a_2}\boldsymbol{X}_i^{\psi} \\ \vdots \\ \mathbb{1}_{a=a_m}\boldsymbol{X}_i^{\psi} \end{pmatrix} w_i \left(Y_i - \boldsymbol{X}^{\beta}\boldsymbol{\beta} - \sum_{\ell=2}^{m} \mathbb{1}_{a=a_\ell}\, \boldsymbol{x}^{\psi}\boldsymbol{\psi}_\ell \right)$$

where $p = q + r(m-1)$ since $\boldsymbol{X}_i^{\beta}$ is $(p \times n)$ and $\boldsymbol{X}_i^{\psi}$ is $r \times n$. While this system of equation may seem intimidating, weighted ordinary least squares can still be used by agglomerating covariates and coefficients from the treatment-free model and blip functions.

Let $\boldsymbol{X}^p = \begin{pmatrix} \boldsymbol{X}_i^{\beta} \\ \mathbb{1}_{a=a_2}\boldsymbol{X}_i^{\psi} \\ \vdots \\ \mathbb{1}_{a=a_m}\boldsymbol{X}_i^{\psi} \end{pmatrix}$ and let $\boldsymbol{\beta}^p = \left(\boldsymbol{\beta}^{\top}, \boldsymbol{\psi}_2^{\top}, \ldots, \boldsymbol{\psi}_m^{\top}\right)^{\top}$ by a $(p \times 1)$ column vector consisting of all treatment-free coefficients and blip function parameters. Using this notation, the score function simplifies to a familiar expression in multiple weighted ordinary least squares. Similarly to the binary case, the optimal treatment can be obtained for patient-specific blip covariates values.

$$\widehat{A}^{\text{opt}}(\boldsymbol{x}^{\psi}) = \begin{cases} \underset{a\in\{1,2,3\}}{\text{argmax}}\ \boldsymbol{x}^{\psi}\widehat{\boldsymbol{\psi}}_a & \text{if } \max\limits_{a\in\{1,2,3\}} \boldsymbol{x}^{\psi}\widehat{\boldsymbol{\psi}}_a > 0 \\ 0 & \text{otherwise} \end{cases}$$

2.5 Longitudinal Framework

In a longitudinal setting, the multi-stage nature of treatment regimes implies that decisions are made in a sequential manner. Ideas in this section – especially the formulation of longitudinal framework for DTRs – are primarily inspired from Chakraborty and Moodie (2013) and Kosorok and Moodie (2015) [5, 21]. When assessing the effects of stage specific treatments, the potential outcome or counterfactual framework is employed in determining the "best case scenario" if all subsequent decisions were optimal. In other words, the stage j counterfactual outcome $\widetilde{Y_j}$ assumes adherence to the optimal regime $\left(a_{j+1}^{\text{opt}}, \ldots, a_K^{\text{opt}}\right)$, which can be estimated by adding the regret functions for stages $j+1, \ldots, K$.

$$\widetilde{Y_j} = Y_j + \sum_{k=j+1}^{K} \mu_k \left(\boldsymbol{x}_k, a_k; \widehat{\boldsymbol{\psi}}_k \right)$$

Finding the optimal sequence of treatments for subsequent stages is possible due to splitting the larger global problem into multiple smaller problems. The backtracking mechanism of the optimization procedure originates from dynamic programming whereby more details about this field are detailed in section 2.5.1. To facilitate the maximization procedure of a final or ultimate outcome in a K-stage decision problem, the value function $V_j^d : \mathcal{H}_j \to \mathbb{R}$ is defined as followed for some stage $j \in \{1, \ldots, K\}$ and for some deterministic or fixed regime $d \in \mathcal{D} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_K$.

$$V_j^d(\boldsymbol{h}_j) = \mathbb{E}\left[\sum_{k=j}^{K} Y_k(\boldsymbol{H}_k, A_k, X_{k+1}) \,\middle|\, \boldsymbol{H}_j = \boldsymbol{h}_j \right]$$

Here, $Y_k(\boldsymbol{H}_k, A_k, X_{k+1})$ represents the kth stage utility or reward given history $\boldsymbol{H}_k$, treatment A_k and subsequent covariates X_{k+1}. Finding the optimal decision vector comprised of K individual actions leading to the best expected terminal outcome value is equivalent to determining the optimal policy d that maximizes V_K^d. Let $V_j^{\text{opt}}(\boldsymbol{h}_j) = \max_{d \in D} V_j^d(\boldsymbol{h}_j)$ denote the maximal expected value of outcome at a possible intermediate stage j. Defining such a function is useful due to breaking down the global optimization procedure into smaller subproblems, which are simpler to solve due to its unidimensionality of action space $\mathcal{A}_j$.

$$\begin{aligned}
V_j^{\text{opt}}(\boldsymbol{h}_j) &= \max_{d \in \mathcal{D}} V_j^d(\boldsymbol{h}_j) \\
&= \max_{d \in \mathcal{D}} \mathbb{E}\left[\sum_{k=j}^{K} Y_k(\boldsymbol{H}_k, A_k, X_{k+1}) \,\middle|\, \boldsymbol{H}_j = \boldsymbol{h}_j \right] \\
&= \max_{d \in \mathcal{D}} \mathbb{E}\left[Y_j(\boldsymbol{H}_j, A_j, X_{j+1}) + \sum_{k=j+1}^{K} Y_k(\boldsymbol{H}_k, A_k, X_{k+1}) \,\middle|\, \boldsymbol{H}_j = \boldsymbol{h}_j \right] \\
&= \max_{d \in \mathcal{D}} \mathbb{E}_{A_j, X_{j+1}}\left[\mathbb{E}_{A_{j+1}, \cdots, A_K, X_{K+1} \mid A_j, X_{j+1}}\left[Y_j(\boldsymbol{H}_j, A_j, X_{j+1}) \right.\right. \\
&\qquad \left.\left. + \sum_{k=j+1}^{K} Y_k(\boldsymbol{H}_k, A_k, X_{k+1}) \,\middle|\, \boldsymbol{H}_j = \boldsymbol{h}_j \right] \right] \\
&= \max_{a_j \in \mathcal{A}_j} \mathbb{E}_{A_j, X_{j+1}}\left[Y_j(\boldsymbol{H}_j, A_j, X_{j+1}) + \max_{d \in \mathcal{D}} \mathbb{E}\left[\sum_{k=j+1}^{K} Y_k(\boldsymbol{H}_k, A_k, X_{k+1}) \,\middle|\, A_j, X_{j+1} \right] \,\middle|\, \boldsymbol{H}_j = \boldsymbol{h}_j \right] \\
&= \max_{a_j \in \mathcal{A}_j} \mathbb{E}_{A_j, X_{j+1}}\left[Y_j(\boldsymbol{H}_j, A_j, X_{j+1}) + V_{j+1}^{\text{opt}}(\boldsymbol{H}_{j+1}) \,\middle|\, \boldsymbol{H}_j = \boldsymbol{h}_j \right]
\end{aligned}$$

For notation simplicity, $\boldsymbol{H}_{j+1}$ was used to denote the accumulated information $(\boldsymbol{h}_j, A_j, X_{j+1})$ even though $\boldsymbol{h}_j$ is given as input in the V_j^{opt} function whereas expectation is taken over A_j and X_{j+1}.

In the context of medical decision making, it is desirable for researchers to perform optimal actions, i.e. ones that maximize the value function. As a result, an analogous function to the one above to address value maximization is the *Q-function* defined as followed at stage j.

$$Q_j^d(\boldsymbol{h}_j, a_j) = \mathbb{E}\left[Y_j(\boldsymbol{H}_j, A_j, X_{j+1}) + V_{j+1}^d\left(\boldsymbol{H}_{j+1}\right) \mid \boldsymbol{H}_j = \boldsymbol{h}_j, A_j = a_j\right]$$

The only difference is that Q_j^d receives the jth stage treatment as input instead of marginalizing over the random variable. Likewise, the optimal value of the *Q-function* at stage j turns out to be the following expression.

$$Q_j^{\text{opt}}(\boldsymbol{h}_j, a_j) = \mathbb{E}\left[Y_j(\boldsymbol{H}_j, A_j, X_{j+1}) + V_{j+1}^{\text{opt}}(\boldsymbol{H}_{j+1}) \,|\, \boldsymbol{H}_j = \boldsymbol{h}_j, A_j = a_j\right]$$

Essentially, the *Q-function* denoted by $Q_j^d(\cdot)$ is almost identical to the valuation function defined earlier, denoted by $V_j^d(\cdot)$. The subtle difference between the two functions is the way A_j is being treated; in $Q_j^d(\cdot)$, A_j serves as an input variable in the conditional expectation whereas, in $V_j^d(\cdot)$, A_j has been marginalized out. In many cases, interest lies in performing inference on the decision space $\mathcal{A}_j$ at stage j and in finding the optimal decision a_j^{opt} that maximizes the utility function, which is why it is often preferable to work with $Q_j^d(\cdot)$ instead of $V_j^d(\cdot)$. Most importantly, the estimation of Q_j^{opt} can be done using linear models.

$$Q_j^{\text{opt}}(\boldsymbol{h}_j, a_j; \boldsymbol{\beta}, \boldsymbol{\psi}) = \boldsymbol{h}_j^{\beta}\boldsymbol{\beta} + a_j \boldsymbol{h}_j^{\psi}\boldsymbol{\psi}$$

As shown in the manipulation of the Q_j^{opt} expression, the optimal jth stage decision requires Q_{j+1}^{opt} to be known and so on. In other words, the optimization of the jth stage treatment variable requires the subsequent ones to be performed first. This backtracking or recursive procedure to global optimization resembles greatly ideas from dynamic programming which are highlighted in section 2.5.1. Likewise, the jth stage model estimates the counterfactual outcome as if all subsequent treatments were chosen to be optimal.

Using the Q_j^{opt} function defined above, the blueprint for regression-based DTR estimation methods for multi-stage problems outlined in Algorithm 1 are in fact similar in design: construction of pseudo-outcome, obtain stage-specific blip coefficient estimates and continue the estimation process for the previous stage. The practical difference between longitudinal version of Q-Learning, G-estimation and dWOLS boils down to a trade-off between theoretical guarantees and difficulty in implementation. In terms of the statistical algorithm itself, the only difference is the score function $U(\boldsymbol{\psi})$ which provides the $\widehat{\boldsymbol{\psi}}$ estimates, as the former varies depending on the estimation method.

2.5.1 Allusions to Other Fields: Dynamic Programming and Reinforcement Learning

The optimization of long-term outcomes through sequential decision making shares common foundations with other quantitative research fields such as dynamic programming and reinforcement learning [3, 47]. The DTR paradigm combines ideas from dynamic programming and statistics in

Algorithm 1: Outline for multi-stage treatment DTR estimation methods

for *j in {K, K-1, . . . , 1}* **do**
  if *j == K* **then**
    Define the pseudo-outcome $\widetilde{\boldsymbol{Y}}_K = \boldsymbol{Y}_K$
  else
    Define the pseudo-outcome $\widetilde{\boldsymbol{Y}}_j = \boldsymbol{Y}_j + \sum_{k=j+1}^{K} \mu_k(\boldsymbol{h}_k, a_k; \widehat{\boldsymbol{\psi}}_k)$
  end
  Set $\mathbb{E}\left[\widetilde{\boldsymbol{Y}}_j \,|\, \boldsymbol{H}_j = \boldsymbol{h}_j, \boldsymbol{A}_j = \boldsymbol{a}_j; \boldsymbol{\beta}_j, \boldsymbol{\psi}_j\right] = \boldsymbol{H}_j^{\beta}\boldsymbol{\beta}_j + \boldsymbol{A}_j\boldsymbol{H}_j^{\psi}\boldsymbol{\psi}$
  Estimate $\widehat{\boldsymbol{\psi}}_j$ in by solving for $\boldsymbol{\psi}$ in a score function $U(\boldsymbol{\psi}_j) = 0$
  Solve for $A_i^{\text{opt}} = \underset{a\in\{0,1\}}{\text{argmax}}\ a\, \boldsymbol{h}_{ij}^{\psi}\boldsymbol{\psi}_j$ for all individuals $1 \leq i \leq n$
end

providing a statistically robust framework in devising the best sequence of decisions. The maximization procedure within DTR paradigm requires that the specified functions $V^d(\cdot)$ and $Q^d(\cdot)$ need to be known in order to maximize them. However, although they are often not known, they can be estimated using data-driven methods. As a result, in finding optimal policies or actions, robust methods are required to posit reliable yet realistic statistical models to estimate the valuation functions.

At its core, as explained above, the DTR optimization procedure in a longitudinal setting solves optimization problems from the end to the beginning. This backtracking process, also known as recursion, is one of the distinguishing feature of dynamic programming which, in short, attempts to optimize multi-stage decision problems by breaking them into multiple intermediate sub-problems. In fact, the DTR estimation procedure adheres to the Bellman equation, which fundamentally originates from dynamic programming, all while estimating model parameters from data [3, 5].

Dynamic programming is a common algorithmic method that solves complex problems recursively by breaking them down into a collection of simpler ones [19]. Reinforcement learning (RL), on the other hand, falls under the umbrella of artificial intelligence and has been an area of research that has received a lot of attention in the past years. In fact, artificial intelligence can be split unto three major subfields: supervised learning, unsupervised learning and RL [13]. In summary, supervised and unsupervised learning deal with extracting information from labelled and unlabelled data respectively, whereas RL focuses on having agents learn how to make optimal decisions when put into different environments. For instance, recent developments in robotics, neuroscience and automated gaming are all products of extensive research in this field of study. RL and DTRs bears many similarities in that they share common methods in tackling problems. In both fields of study, many problems are formulated to maximize some reward through sequential decision making. More formally put, the RL framework postulates that an agent is presented a set of actions $\mathcal{A}$ while being put in some environment or state $s \in \mathcal{S}$. The agent's probability in selecting a given action or decision $a \in \mathcal{A}$ for

some state s is represented by $\pi(a, s) \in (0, 1)$ known as a policy map.

$$\pi(a, s) : \mathcal{A} \times \mathcal{S} \to [0, 1]$$
$$R(a, s) : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$$

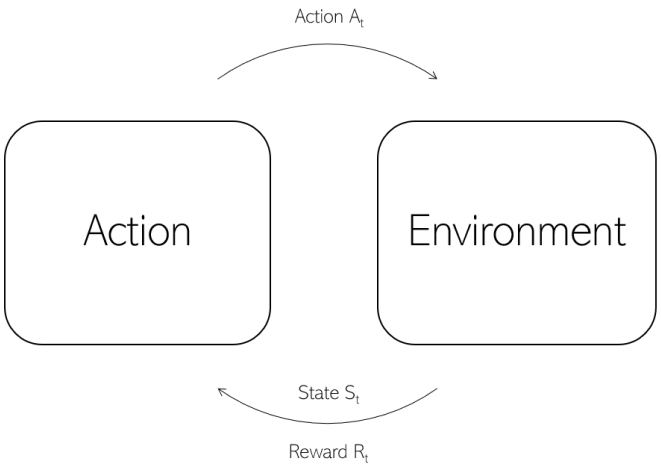

Figure 1: Visual representation of iterative process of the decision making coupled with the generation of states and rewards [47].

The iterative process of changing states, obtaining rewards and forming decisions in discrete time is known as a Markov decision process [2, 47]. Notice how the formulation of problems in reinforcement learning bears strong similarities with DTRs except with slightly different terminologies. Commonly used terms such as rewards, actions, states and policy maps are analogous to outcomes, treatments, covariate information or patient history and propensity scores when talking about DTRs.

Notable differences between these two fields are primarily the complexity of problems which are being tackled and the methods which are used to solve them. Finite-horizon problems are ones where a fixed number of stages or a stopping rule is posited in preventing the problem from continuing endlessly. The lack of a foreseeable end in such problems often entail extremely complex environments (such as in robotics, locomotion and automated vehicle control) and novel yet relatively computationally intensive estimation methods with fewer known theoretical properties [20, 10, 35, 49]. In DTRs, while problems are often of lower complexity, methods in solving for optimal decisions are rather more interpretable and easily implementable at the exchange of the single structure in the potential studied problems. However, recent works have explored the usage of more flexible high-dimensional techniques in estimating treatment strategies [25].

2.6 Data Analysis

2.6.1 Preliminary Definitions

We begin with definitions, as there are many terms that are similar or make have similar non-technical meanings.

Definition 2.9 (Injection cycle)**.** An **injection cycle** or cycle of injections refers to a group of repeated injections, each of which were administered at one-week intervals. Each cycle is comprised of $1, 2$ or 3 injections, all of which were administered at a dosage of $20\mu\text{g/kg}$.

Definition 2.10 (Treatment interval)**.** A **treatment interval** is a period which spans from the beginning of an injection cycle to the beginning of the subsequent injection cycle, if there are any, or to the end of the patient's involvement in the clinical trial. For participants in the control arm of INSPIRE 2, the beginning of their participation to the beginning of their first injection cycle, which spans approximately one year, is considered a treatment interval with no injection cycles. For all other cases, a treatment interval is specified by a series of non-zero repeated injections.

Definition 2.11 (Treatment stage)**.** A **treatment stage** is an approximate 90-day period whereby patients are eligible to receive a cycle of repeated injections[6]. Each treatment interval can be divided into one or multiple treatment stages. In the case where a treatment interval consists of more than one treatment stage, the first stage is where the administration of injections take place and the subsequent ones are treatment stages with 0 injections.

Definition 2.12 ($CD4_{\text{init}}$)**.** The initial concentration at the beginning of a treatment stage denoted by $CD4_{\text{init}}$ is the CD4 load of the first measurement of the stage.

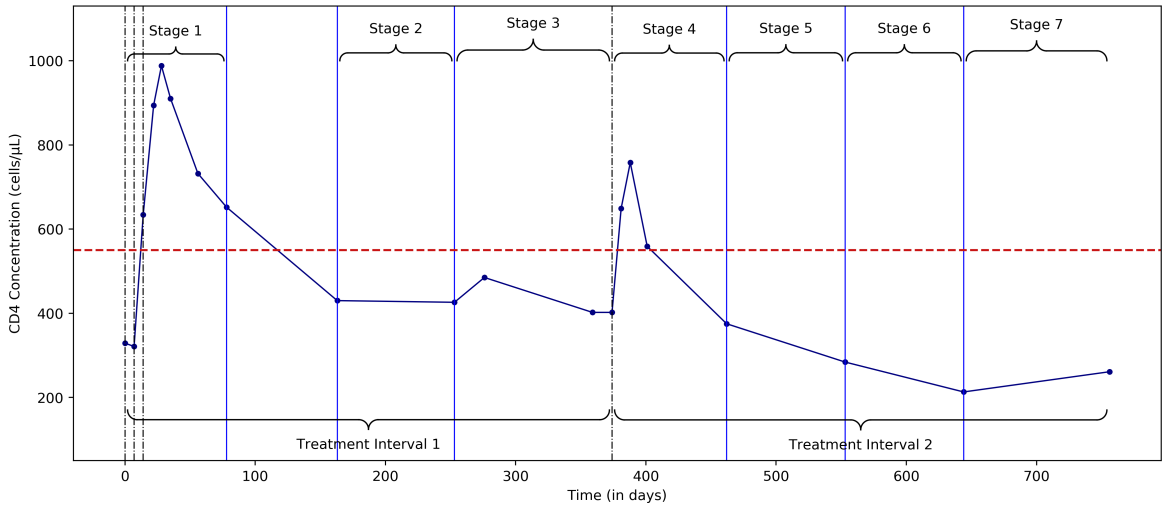

Figure 2: CD4 dynamics of patient 1701 from INSPIRE 2.

[6]Recall that patients need to have CD4 load less than 550 cells/μL at the beginning of the treatment stage to receive injections.

Example 2.1. Consider the CD4 dynamics of patient 1701 with the labelling of treatment stages and treatment intervals in Figure 2. In this illustrative example[7], this participant has received two injection cycles: a first cycle consisting of 3 injections spanning from day 0 to day 374 and a second cycle consisting of a single injection spanning from day 374 to day 756. Both treatment intervals are split into multiple treatment stages: the first treatment interval contains 3 treatment stages whereas the second cycle contains 4 treatment stages. In the first treatment interval, the participant receives an injection cycle consisting of 3 injections in the first treatment stage and this is followed by 3 treatment stages where no injections are administered. In the second treatment interval, the participant receives an injection cycle consisting of 1 injection in the first treatment stage (4th overall) and this is followed by 3 treatment stages with no injections. Notice that the interval between day 78 and day 163 does not constitute as a treatment stage because the CD4 count measured at day 78 is above 550 cells/μL. Generally speaking, treatment interval is a term which is more relevant when talking about the clinical trial whereas treatment stage is a term that is better suited for statistical analysis.

2.6.2 Data Adaptation for DTR Framework

In ITR estimation, focus lies in maximizing a "good outcome" that is reflective of a patient's well-being. Because the research question involves assessing the proportion of time where the CD4 concentration is above 500 cells/μL over the short-term duration following a treatment decision, the first challenge was to adapt the data at hand to conform to a DTR question. To do so, we conceived of a scalar-valued outcome consisting of a utility accounting for a patient's immune response and penalization for excessive injections, so that the outcome of interest was a single measurement representing an entire 90-day treatment stage. Every observation in the processed data consists of patient-stages where $CD4_{\text{init}} < 550$ cells/μL.

An outcome of interest denoted by U^g is defined to be the proportion of time, following treatment, that a participant had a healthy CD4 load, capturing immune response over a given treatment stage. With multiple measurements available for each treatment interval, a patient's CD4 dynamics need to be estimated. While more sophisticated statistical methods such as mechanistic models [18] have been used to estimate the T cell dynamics, for simplicity we use linear interpolation to estimate the trajectory of patients cell count.

Definition 2.13 ($CD4(t)$)**.** Given a set of n observations $\{(t_k, CD4_k)\}_{k=1}^n$ over some treatment stage, the trajectory $CD4(t)$ can be estimated over $[t_1, t_n]$ as a piecewise function comprised of linear functions.

$$CD4(t) = \frac{CD4_{k+1} - CD4_k}{t_{k+1} - t_k}(t - t_k) + CD4_k \qquad t \in [t_k, t_{k+1}] \text{ for all } k \in \{1, \ldots, n-1\}.$$

In subsequent sections, it could be of particular interest to investigate the first measurement taken at the beginning of a treatment stage; this specific $CD4$ cell count will be represented by the notation $CD4_{\text{init}}$.

[7] In the graph, every "stage" is short for "treatment stage".

Definition 2.14 (U^g)**.** The utility U^g is defined to be the **proportion** of time where the patient of interest has a CD4 concentration greater or equal to 550 cell/μL. Formally, given a trajectory $CD4(t)$ defined over some interval $[t_1, t_n]$, U^g is defined as follows.

$$U^g = \frac{1}{t_n - t_1} \int_{t_1}^{t_n} \mathbb{1}_{CD4(t) \geq 550} \, \mathrm{d}t$$

For instance, by definition, $U^g \in [0, 1]$, where $U^g = 0$ if all observed measurements are below 550 cells/μL.

Definition 2.15 (U^i)**.** The cost U^i is defined to be the negative of the number of injections. In other words, the purpose of penalizing for the number of injections on top of a low CD4 cell count is to prevent the administration of superfluous injections. Thus, $U^i \in \{-3, -2, -1, 0\}$.

Example 2.2. Consider the following illustrative example of a patients' CD4 dynamics and two injections were received in this treatment stage and are represented by the two vertical dotted lines. The three CD4 measurements are $400, 700$ and 500 taken at respectively day $0, 24$ and 80.

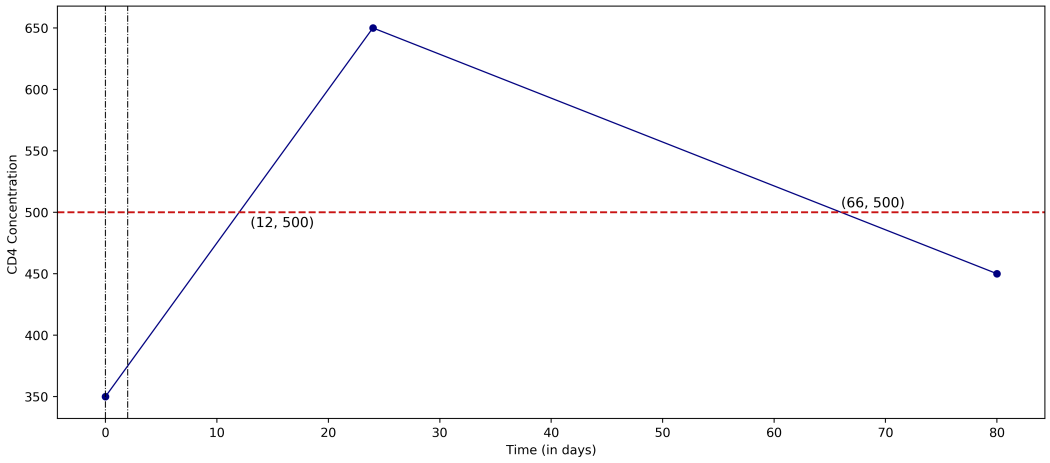

Figure 3: Estimation of CD4 dynamics using linear interpolation.

Because two injections were administered, $U^i = -2$. The estimated CD4 trajectory in Figure 3 crosses the 500 cells/μL threshold at time points $t \in \{12, 66\}$. It follows that $CD4(t) \geq 0$ for $t \in [12, 66]$ and the utility associated to immune response can be calculated as followed.

$$\begin{aligned} U^g &= \frac{1}{80} \int_0^{80} \mathbb{1}_{CD4(t) \geq 550} \, \mathrm{d}t \\ &= \frac{1}{80} \int_{12}^{66} 1 \, \mathrm{d}t \\ &= 0.55 \end{aligned}$$

Definition 2.16 ($U^\eta(\eta)$)**.** Inspired by Pasin et al., the outcome variable, denoted by $U^\eta(\eta)$, is defined to be a convex sum of two previously defined utilities, U^g and U^i.

$$U^{\eta}(\eta) = \eta U^{g} + (1-\eta)U^{i} \qquad \text{for } \eta \in [0,1]$$

The parameter η allows the focus of the constructed outcome variable to vary depending on the chosen value of η. Thus, an η of 0 would suggest that the utility is simply the negative of the number of injections, which would be maximized by never injecting participants. Conversely, setting η to 1 would generate a treatment rule designed to maximize CD4 response, without any consideration of the number of injections.

2.6.3 Tailoring Variables

Personalized data collected in the INSPIRE studies 2 and 3 can be used to tailor the number of injections to HIV-infected patients. In the outcome model, such quantities can be incorporated as part of the blip function and they include the following information: age, sex, BMI and ethnicity. A table summarizing the details of these quantities, such as mean and standard deviation, is given in Table 2.

In practice, the idea behind longitudinal studies as a whole is to investigate the effects of studied interventions over time. The repeated assessment of outcome measures allows researchers to observe the trajectory or changes in a participant's well-being, and leverage the repeated measures by treating individuals as their own controls by comparing different treatments given to the same individual at different times. To do so, studies typically span a sufficiently long period of time to allow any effects of medical interventions to develop and to be observable. Looking at myopic regimes implies that the sequential nature of treatment stages is omitted in the analysis, and our focus is limited to short-term changes in outcome effected by the treatment. We allow for previous immune responses to affect future ones in the modelling procedure by conditioning on individual patients' histories. That is, in addition to positing a correlation structure between treatment stages of a same patient, one way to account for historical injection information is to define tailoring variables that can embody potential prior biological response to the administration of IL-7. Variables defined below and denoted by Hx and Resp are used to portray historical treatment information that can be relevant in tailoring future treatments.

Definition 2.17 (Hx)**.** A patient's historical treatment information, denoted by Hx, is a dichotomous variable indicating if a patient has received injections in a prior treatment stage:

$$\text{Hx} = \begin{cases} 0 & \text{if no injections were previously administered} \\ 1 & \text{if at least 1 injection was previously administered.} \end{cases}$$

By definition, for any patient, treatment stages will have an Hx value of 0 until an injection is administered. As soon as a treatment stage is observed in which a participant receives at least 1 injection, the Hx covariate will take a value of 1 for all subsequent treatment stages for that patient.

Definition 2.18 (Resp)**.** A patient's response to previous treatment is denoted by Resp, and is defined

as

$$\text{Resp} = \begin{cases} 0 & \text{if Hx } = 0 \\ \frac{1}{n^{\text{prev}}} \left(\max_k \left(CD4_k^{\text{prev}} \right) - CD4_{\text{init}}^{\text{prev}} \right) & \text{if Hx } = 1. \end{cases}$$

where the prev subscript refers to the preceding treatment stage where a cycle of injections was provided to the individual. In the definition of Resp, the multiplicative coefficient $\{n^{\text{prev}}\}^{-1}$ primarily serves the purpose of adjusting for the number of injections administered. In essence, the idea is to provide a measure in CD4 "jump" attributable for a single injection. For instance, for a same increase in CD4 count, the Resp variable should exhibit a larger value if fewer injections were received. The idea is to highlight the sensitivity participants' immune response to the quantity of IL-7 provided.

Example 2.3. Refer back to the dynamic of patient 1701 displayed in Figure 2. The first cycle of injections are administered in the first stage; as a result, Hx $= 0$ in stage 1 but Hx $= 1$ for all subsequent stages, i.e. for stages $2, 3, \ldots, 7$. Likewise, Resp $= 0$ for stage 1. For stage $2, 3$ and 4, the most recent treatment stage with a non-zero number of injections was stage 1. As a result, they bear the same Resp value of $\frac{1}{3}(988 - 329) \approx 219.7$; stages $5, 6$ and 7 all have a Resp value of 356, which was obtained using the same calculation on treatment stage 4.

In the dWOLS analysis, the natural logarithm of the above-defined response shifted by 1 unit, denoted logResp, will be used in the analysis:

$$\text{logResp} = \log(\text{Resp} + 1).$$

Resp exhibited considerable skew, and so this transformation was employed to reduce the impact of outlying values.

2.7 Analysis Plan

The statistical analysis of INSPIRE data using an ITR framework is summarized in the following protocol.

1. Organize the data into a long format and compute U^g and U^i for all patients and all stages. For instance, the graph above displays the CD4 cell count concentration over time for patient 1701 in Figure 2. Merging baseline and time-varying information yields Table 1, whereby data is organized in a long format and ready for dWOLS implementation.
   In the data collected during the INSPIRE studies, there are three categories for Origin: Caucasian, African and Other; however due to small numbers, this is recoded as a binary variable, where 1 encodes a participant being of African origin and 0 otherwise (i.e. Caucasian or Other). The Sex variable was also defined as dichotomous, where 1 represents male.
2. Fit multinomial propensity scores $P(A_i = a_i \,|\, \boldsymbol{X}_i^\alpha)$ where covariates $\boldsymbol{X}_i^\alpha$ consist of the following: Sex, Age, BMI, Origin and $CD4_{\text{init}}$. Weights w_i can be taken to be $P(A_i = a_i \,|\, \boldsymbol{X}_i^\alpha)^{-1}$. We use $A = 0$ as baseline treatment as it is the largest group.

Table 1: Patient 1701 data in "long" format, obtained by extracting and combining relevant information from baseline information, observed time-varying information, and outcome data generated through the linear interpolation of CD4 counts.

Patient ID	Sex	Age	BMI	Origin	Hx	Resp	$CD4_{\text{init}}$	U^g	Number of Injections (U^i)
1701	0	49.9	26.48	0	0	0	329	0.859	3
1701	0	49.9	26.48	0	1	219.7	430	0	0
1701	0	49.9	26.48	0	1	219.7	426	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3. Compute $U^\eta(\eta) = \eta U^g + (1-\eta)U^i$ for a given weight value η for all individual-stage data. Separate analyses will be performed for each utility.
4. Define the treatment-free and blip model covariates to include patient-specific data:
   - $\boldsymbol{X}^\beta$: Sex, Age, BMI, Origin, Hx, logResp and $CD4_{\text{init}}$
   - $\boldsymbol{X}^\psi$: Sex, Age, BMI, Origin, Hx and logResp
5. Apply the categorical dWOLS algorithm above using the defined IPTW weights W, treatment-free covariates $\boldsymbol{X}^\beta$, blip covariates $\boldsymbol{X}^\psi$ and outcome variable $U^\eta(0)$. We use ψ_a for the vector of blip parameters associated with injections for $a \in \{1,2,3\}$ (where, recall, $A=0$ is the baseline treatment category.
6. Estimate empirical standard errors and confidence intervals using either the robust variance estimator or the bootstrap procedure for the blip parameters $\widehat{\boldsymbol{\psi}}$ obtained from the previous step.
7. Determine the optimal treatment (number of injections) for each patient using the formula provided at the end of section 2.4.3.

Repeat the analytic steps 5-7 for other values $\eta \in [0,1]$, first recomputing the utility in step 3.

3 Results

3.1 Descriptive Results

A summary of covariates associated to each patient-stage in the processed data is available in Table 2. The standardized mean difference (SMD) is a score that measures the imbalance of characteristics across observations in different treatment groups [11, 55]. An SMD value greater than 0.1 is a common criterion used to determine if a particular covariate is imbalanced across treatment groups [44]. For instance, in Table 2, both the Age and Sex variable have an SMD value of 0.12, the smallest SMD value amongst covariates of interest. This descriptive measure shows that there is significant imbalance in all characteristics across observations grouped by treatment category. It is also worth highlighting the low sample size in treatment categories $A = 1$ and $A = 2$. There are 315 observations associated with 0 injections and 150 observations associated with 3 injections whereas there are only 17 and 22 observations associated with 1 and 2 injections respectively. The low number of observations in the $A = 1$ and $A = 2$ group is a by-product of the INSPIRE protocols where 3 injections should be have administered according to study protocol.

Table 2: Summary of patient characteristics with respect to number of injections received

Characteristic	Mean (SD)				SMD
	$A = 0$ ($n = 315$)	$A = 1$ ($n = 17$)	$A = 2$ ($n = 22$)	$A = 3$ ($n = 150$)	
Sex	0.77 (0.44)	0.71 (0.47)	0.77 (0.43)	0.68 (0.47)	0.12
Origin	0.40 (0.49)	0.29 (0.47)	0.32 (0.48)	0.48 (0.50)	0.22
BMI	24.4 (3.6)	25.0 (4.5)	25.8 (4.5)	24.3 (3.5)	0.21
Age	45.4 (8.8)	46.3 (8.8)	44.5 (7.9)	44.9 (8.4)	0.12
$CD4_{\text{init}}$	350 (112)	435 (89)	358 (105)	322 (116)	0.56
logResp	3.60 (2.53)	4.89 (1.90)	2.97 (2.79)	2.01 (2.68)	0.64

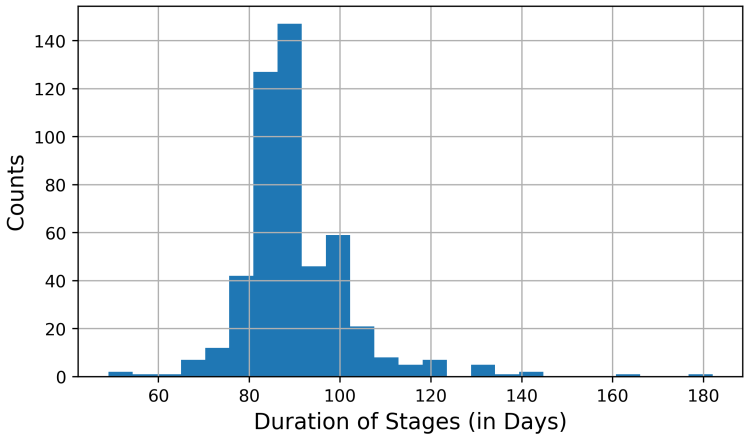

Figure 4: Histogram of durations of all eligible stages

The distribution of the treatment stage duration in days is displayed in Figure 4. Because the research question examines the ideal number of injections in a cycle for each 90-day window, the distribution is tightly centred near 90 days. The observed heterogeneity in interval durations is most likely attributable to randomness in appointment scheduling during the clinical trials. It is assumed that this variability does not substantially affect the analysis, especially in the definition of the immune response utility, where U^g is computed as a proportion.

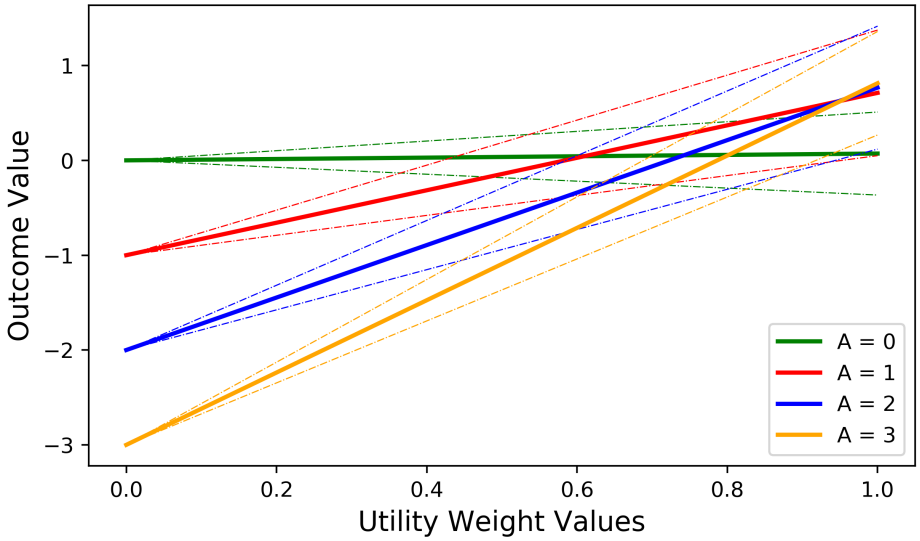

Figure 5: Average $U^{\eta}(\eta)$ values with 95% confidence intervals with respect to treatment group plotted across η values in $[0, 1]$.

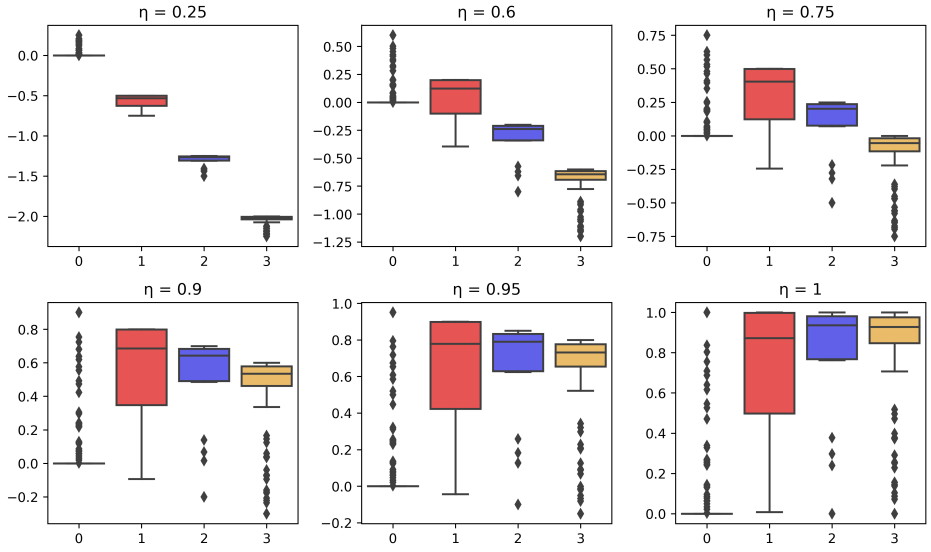

Figure 6: Boxplots for $U^{\eta}(\eta)$ values for $\eta \in \{0.25, 0.6, 0.75, 0.9, 0.95, 1\}$.

The utility (outcome), averaged over all individual treatment stages in the study, are given under each number of injections in Figure 5: the average $U^{\eta}(\cdot)$ value across observations and their 95% confidence intervals are represented by a thick line and dotted lines, respectively. The linearity of means and confidence interval bounds follows immediately from $U^{\eta}(\cdot)$ being a linear combination of U^{g} and U^{i}. Boxplots for $U^{\eta}(\eta)$ are displayed in Figure 6, stratified by the number of injections provided in patient-stages for several η values. Together, these two figures provide an overview of the outcome measure distribution with respect to treatment categories and utility weights. For instance, for $\eta < 0.6$, the value of $U^{\eta}(\cdot)$ is largely dominated by the penalizing term U^{i} for the quantity of administered injections. This result follows from the discrepancy in the widths of ranges of U^{g} and U^{i}; $U^{g} \in [0, 1]$ whereas $U^{i} \in \{-3, -2, -1, 0\}$. When more emphasis is put on immune response, i.e. in cases where η conveys a larger value such as $0.9, 0.95$ or 1, the outcome values for $1, 2$ and 3 injections are closer to each other. In the bottom three plots in Figure 6, utilities capturing immune response information do not seem to be considerably different in patient-stages where at least 1 injection was administered. U^{g} values for 0 injections are 0 for nearly all patient stages, which explains the lack of variability in the boxplot for $A = 0$.

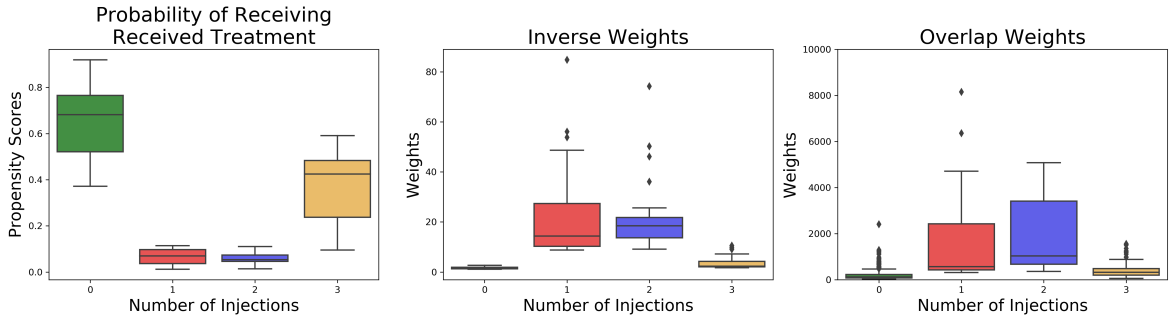

Figure 7: Histogram of propensity scores, inverse weights and overlap weights

Generalized propensity scores were fitted using multinomial logistic regression, which were then used to compute the weights w_i needed for the dWOLS analysis. Boxplots for the propensity score $P(A = a \,|\, \boldsymbol{X})$, inverse probability treatment weights and overlap weights are displayed in Figure 7. The boxplots of the generalized propensity scores show that, in general, participants have a higher likelihood to receive 3 or no injections compared to receiving 1 or 2 injections in an injection cycle. This follows immediately from the small sample size of patient-stages where only 1 or 2 injections were administered as a cycle. According to middle and right-hand panel, titled "Inverse weights" and "Overlap weights", it is shown that observations associated with treatment groups $A = 1$ or $A = 2$ receive larger weight values in the dWOLS analysis to accommodate for their underrepresentation in the dataset.

3.2 Statistical Summary of Blip Coefficients

Estimates of blip parameters $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ are dependent on the outcome measure $U^{\eta}(\cdot)$ which itself relies on a particular choice of $\eta \in [0, 1]$. Each η value yields different values for the outcome variable, which in turn implies that the estimated ITR would vary accordingly. Because there are many coefficient estimates of interest and the utility weight is a continuous parameter, the overarching purpose of the statistical analysis and its interpretation is to paint a comprehensive portrait of the results while providing sufficient detail to determine key factors that influence treatment recommendation. The statistical analysis results for $\eta = 0.7$ and 0.9 are provided in Tables 3 and 4 respectively on the following page. Coefficient estimates are displayed alongside their 95% confidence intervals; bold estimates are statistically significant at a 5% level.

From these summary tables, coefficients for Hx and Resp, denoted by $\widehat{\psi}_{\ell,\text{Hx}}$ and $\widehat{\psi}_{\ell,\text{logResp}}$, are statistically significant across contrast functions $\gamma_\ell(\cdot)$, $\ell = 1, 2, 3$ for both $\eta = 0.7$ eta $\eta = 0.9$. Other coefficients such as $\widehat{\psi}_{2,\text{Sex}}, \widehat{\psi}_{3,\text{Sex}}$ and $\widehat{\psi}_{3,\text{Age}}$ are also statistically significant for $\eta = 0.7$ and $\eta = 0.9$. Recall that, by construction of the Hx and Resp variables representing IL-7 injection history, patients with a non-zero logResp value have an Hx value of 1. Consider the following expression from the blip function capturing information attributable to prior injections.

$$\widehat{\psi}_{\ell,\text{Hx}}\text{Hx} + \widehat{\psi}_{\ell,\text{logResp}}\text{logResp} \tag{1}$$

Table 3: Summary of estimated blip coefficients for a dWOLS analysis of outcome $U^{\eta}(0.7)$

Characteristic	Estimates (95% C.I.)					
	$A = 1$		$A = 2$		$A = 3$	
Intercept	-0.11	$(-0.64, 0.41)$	-0.10	$(-0.83, 0.63)$	0.17	$(-0.11, 0.45)$
Age[a]	0.08	$(-0.01, 0.17)$	-0.07	$(-0.18, 0.04)$	**-0.04**	$(-0.07, 0.00)$
Sex	-0.02	$(-0.11, 0.06)$	**0.22**	$(0.00, 0.43)$	**-0.06**	$(-0.12, 0.00)$
BMI[b]	-0.03	$(-0.16, 0.10)$	0.03	$(-0.13, 0.19)$	**-0.12**	$(-0.21, -0.04)$
Origin	0.14	$(-0.02, 0.29)$	0.11	$(-0.08, 0.30)$	-0.02	$(-0.08, 0.05)$
Hx	**-1.79**	$(-2.69, -0.89)$	**-1.35**	$(-2.15, -0.55)$	**-0.65**	$(-1.08, -0.21)$
logResp	**0.29**	$(0.13, 0.45)$	**0.24**	$(0.10, 0.39)$	**0.10**	$(0.03, 0.18)$

Table 4: Summary of estimated blip coefficients for a dWOLS analysis of outcome $U^{\eta}(0.9)$

Characteristic	Estimates (95% C.I.)					
	$A = 1$		$A = 2$		$A = 3$	
Intercept	0.14	$(-0.54, -0.82)$	0.44	$(-0.50, 1.4)$	**1.08**	$(0.72, 1.44)$
Age[a]	0.10	$(-0.01, 0.21)$	-0.09	$(-0.23, 0.06)$	**-0.05**	$(-0.10, 0.00)$
Sex	-0.03	$(-0.14, 0.08)$	**0.28**	$(0.04, 0.56)$	**-0.08**	$(-0.20, 0.00)$
BMI[b]	-0.04	$(-0.20, 0.13)$	0.04	$(-0.17, 0.25)$	**-0.16**	$(-0.28, -0.05)$
Origin	0.18	$(-0.03, 0.38)$	0.14	$(-0.11, 0.38)$	-0.02	$(-0.11, 0.06)$
Hx	**-2.30**	$(-3.45, -1.14)$	**-1.73**	$(-2.77, -0.70)$	**-0.83**	$(-1.39, -0.27)$
logResp	**0.37**	$(0.17, 0.57)$	**0.31**	$(0.12, 0.50)$	**0.13**	$(0.033, 0.23)$

[a]For every 10 years
[b]For every 10kg/m^2

Although $\widehat{\psi}_{\ell,\text{logResp}}$ bears a positive value for $\ell = 1, 2, 3$ and $\eta = 0.7, 0.9$, estimates for $\widehat{\psi_{\ell,\text{Hx}}}$ are negative. By plugging in various logResp values, it can be shown that the equation (1) is more likely to be positive when participants exhibit a large logResp value. This in turn can imply one of two things: either response to previous IL-7 injections was good or they began the treatment stage with a very low CD4 count. For instance, say that, for illustrative purposes, three people have Resp values of $200, 500$ and 700. The evaluation of the expression in (1) for $\ell = 1, 2, 3$ and for $\eta = 0.7, 0.9$ is summarized in the Table 5.

logResp values	$\ell = 1$	$\ell = 2$	$\ell = 3$
200	-0.262	-0.027	-0.120
500	0.003	0.192	-0.028
700	0.100	0.273	0.005

logResp values	$\ell = 1$	$\ell = 2$	$\ell = 3$
200	-0.338	-0.056	-0.141
500	0.000	0.227	-0.022
700	0.124	0.331	0.022

Table 5: Evaluation of $\widehat{\psi}_{\ell,\text{Hx}}\text{Hx} + \widehat{\psi}_{\ell,\text{logResp}}\text{logResp}$ for different treatment options $\ell = 1, 2, 3$ fixing $\eta = 0.7$ and $\eta = 0.9$.

From this, it can be inferred that the estimated ITRs for $\eta = 0.7, 0.9$ generally recommend the administration of a cycle of injections if benefits of previous injections have been reflected in an increase of CD4 cell count, information which is conveyed through the definition of the Resp variable. Although the parameter estimates for lower η values are not displayed in this section, it is the case that results suggest a conservative approach when recommending injections.

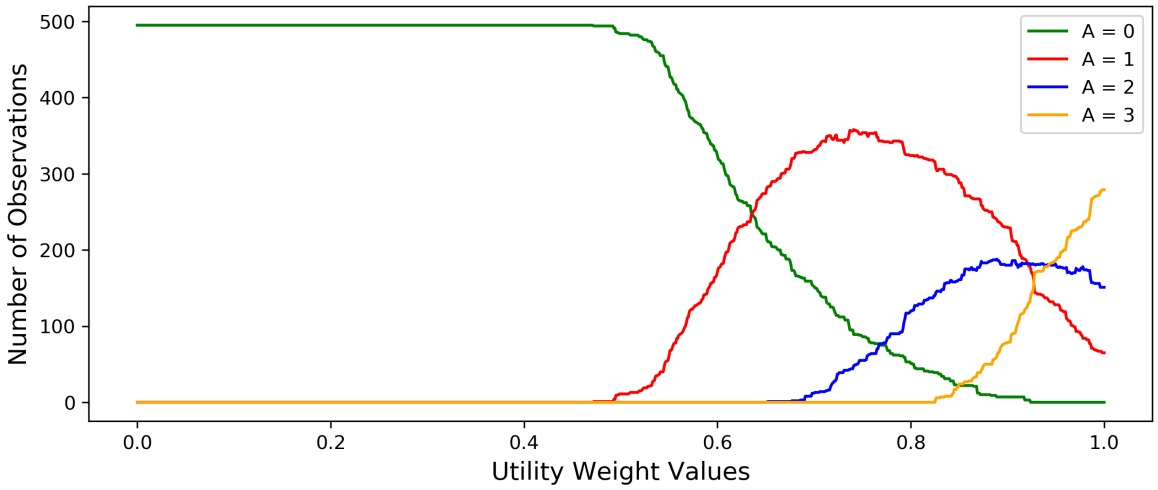

Figure 8: Number of observations having A^{opt} as each treatment type with respect to η values.

The optimal number injections can also be calculated for each patient-stage in the dataset used for the analysis. Estimation of contrast values $\gamma_\ell(\boldsymbol{h}^\psi)$ can be done by plugging in covariate-specific information $\boldsymbol{h}^\psi$ for each patient-stage in the processed dataset. $\widehat{A}^{\text{opt}}$ can be obtained by comparing values of the blip functions as discussed in section 2.4.2. The number of observations having $\widehat{A}^{\text{opt}}$

being equal to $0, 1, 2$ or 3 injections are displayed with respect to η values ranging from 0 to 1 in Figure 8. According to this graph, for smaller values of η, i.e. approximately when $\eta < 0.4$, the recommended number of injections is 0, which follows immediately from most of the utility weight being assigned to minimize the number of injections. As η increases its value beyond the 0.4 mark, more patient-stages are recommended to inject once, which can be inferred from the gradual increase in the red curve. Beyond the $\eta = 0.6$ mark, a cycle consisting of 2 injections is more likely to be recommended and likewise for 3 injections when η increases beyond the 0.8 threshold. Having a treatment rule that suggests more injections for values of η closer to 1 is due to a larger importance attributed to a stronger immune response rather. One important thing to observe is that, when $\eta = 1$, i.e. when there is no penalization for the number of injections, a considerable number of patient-stages in are still being recommended 1 or 2 injections rather than 3. Although other factors such as a patient's age, sex, ethnic origin may affect the estimated ITR, recall that the observed immune response utility U^g seems to be comparable across groups $A = 1$, $A = 2$ and $A = 3$ (c.f. Figure 6).

3.3 Treatment Recommendation for Specific Patient Profiles

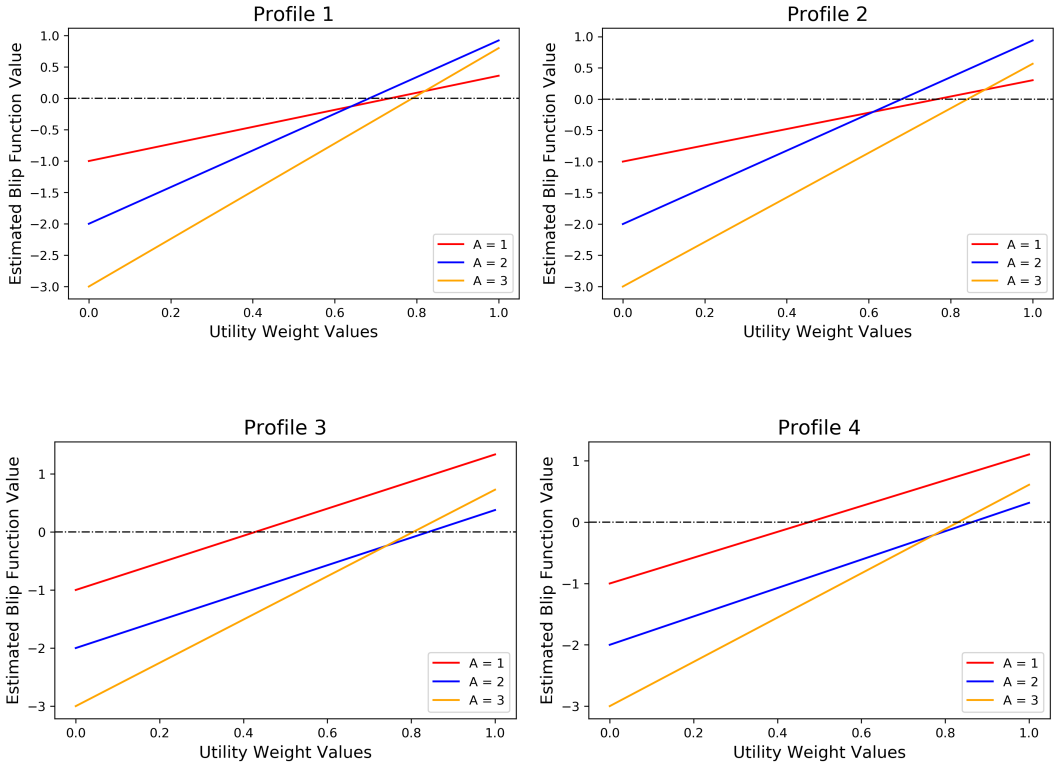

Figure 9: Contrast function utility for different number of injections with respect to utility weights $\eta \in [0, 1]$ for four patient profiles.

Treatments depend on the chosen weight values associated to utilities U^g and U^i; however the individualization of the treatment recommendation also depends on characteristics of patients, as the blip model is a function of covariates $\boldsymbol{X}^{\psi}$. To give an idea of the optimal number of injections, plots

showing the estimated value of $\gamma_\ell(\cdot, a_\ell)$ for $a_\ell = 1, 2, 3$ for four chosen patient profiles are shown in Figure 3.3. The chosen patient profiles are as follow:

Profile 1: 25-year-old man of non-African ethnic background, with a BMI of 25 and without any IL-7 injection history;

Profile 2: 40-year-old woman of African ethnic background, with a BMI of 35 and with a recent injection cycle with a Resp value of 200;

Profile 3: 60-year-old man of African ethnic background, with a BMI of 24 and without any IL-7 injection history;

Profile 4: 80-year-old woman of non-African ethnic background, with a BMI of 30 and with a recent injection cycle with a Resp value of 400.

The dotted line representing the 0 utility threshold is the value of the blip function for the $A = 0$ category since it was chosen to be reference treatment. Because the blip function compares utilities across treatment options, the number of injections exhibiting the largest utility value for a fixed η value is the optimal treatment determined by the dWOLS analysis. For instance, for patient profile 1, for values of $\eta < 0.6$, the results suggest that no injections should be administered. On the other hand, for values larger than approximately 0.7, a cycle of 2 injections is recommended. From two plots, three main conclusions can be drawn.

Firstly, similarly to what Jarne et al. have found [18], there does not seem to be a considerable benefit in administering 3 injections. Instead, either 1 or 2 injections seem to result in comparable or even better immune response since, for $\eta = 1$, blue curves are above yellow ones for patient profiles 1 and 2 whereas red curves are above yellow ones for patient profiles 3 and 4. Secondly, while it is intuitive that $A^{\text{opt}} = 0$ for smaller values of η, the decision to administer a non-zero number of injections seems to vary across patient profiles. The red line crosses the dotted line in profile 3 and 4 for a smaller value of η relative to the blue line in profiles 1 and 2. The main difference between these pairs of patient profiles is their age: profiles 1 and 2 are respectively 20 and 40 years old whereas profiles 3 and 4 are respectively 60 and 80 years old. This in turn can be imply that, according to this analysis, a first IL-7 injection is more necessary in older participants than in younger counterparts but a second dosage is unnecessary. However, for younger patients, assigning $1, 2$ or 3 injections yield comparable results in immune response; a moderate decision of 2 injections may be the best, most conservative approach.

4 Discussion and Conclusion

Overall, the dWOLS analysis provides valuable insight on the ideal number of injections through the design of ITRs, one for each value of the utility weight η. Treatment recommendation largely depends on two things: the value of η and patient-specific information. When η is closer to 0, more weight is put on minimizing the number of injections and, as a result, the recommended number of injections would be more conservative. When η is closer to 1, more important is put on having a better immune response, hence participants are more likely to be recommended to receive injections. However, despite this, many observations in the dataset are suggested to receive 1 or 2 injections. This result follows immediately from the good immune response outcome for observations associated to these treatment categories, despite their low sample size (see Figure 6). A relatively large number of treatment stages are not being recommended to administer the maximal number of injections set by the clinical protocol, which warrants further investigation on the relevance of a third injection first raised by Jarne et al. [18] While a third injection may increase the CD4 load to a higher peak, their results show that two injections are sufficient in ensuring that the CD4 concentration goes above the threshold of 500 cells/μL [18, 34].

Previous approaches to optimizing IL-7 treatment protocols warranted investigation on the magnitude of the effect of other covariate-specific information on the ideal number of injections to provide. In this work, intrinsic patient characteristics such as sex, BMI and ethnic origin have not been shown to be statistically and clinically significant in individualizing the number of injections. Although our analyses were limited in power, our results suggest that age may be clinically useful tailoring variable to determine the optimal number of injections. For instance, the findings in the analysis of specific patient profiles recommend a first injection in older patients at lower utility weights than in younger patients. However, for the latter subpopulation, a second injection seems more necessary whereas, in older patients, treatment recommendation seem to follow a "one or nothing" approach to injections. This raises awareness to the treatment toxicity in older PLWH and especially the ability of their immune system to handle the injections. On the other hand, this also highlights the potency of such treatment in younger participants, since the benefits of a single injection are outweighed by its risk and the benefits of two injections (see Figure 3.3). The effect of age on treatment recommendation warrants further investigation regarding this finding, since adverse effects to IL-7 were not recorded in the data.

Two major limitations that are worth addressing are the sparsity of observations associated with 1 and 2 injections and the design of the myopic rule. When receiving a cycle of IL-7 injections, participants from the INSPIRE studies were more much more likely to receive 3 injections than they are to receive 1 or 2; cycles consisting of 1 or 2 injections are considered incomplete by the clinical protocols of INSPIRE 2 and 3 [51, 53]. The design of the studied ITR assumes that the effect of IL-7 injections are short-term, in that their lingering influence on the CD4 count across subsequent stages are insignificant. The use of an ITR to investigate injection effects is motivated by the desire to maximize immediate utilities and its simpler interpretation [36]. An observable trend in the INSPIRE participants' CD4 dynamics is its gradual decrease over time following injections. An important model-based consideration would be the duration since prior injections because, with increasing

time, patients CD4 load are be more likely to fall below 500 cells/μL. Although the logResp variable attempted to capture information regarding immune response from previous injections, the temporal aspect was not considered in this analysis.

5 Experience as a Professional

Working on this project has been a great experience for me. Not only was I able to acquire valuable research experience in two countries, I had the opportunity to meet many individuals of similar and different background at various stages in their career paths. The research team at the University of Bordeaux had more experience in working with mechanistic models whose parameters are estimated using a Bayesian approach. The research focus in the biostatistics community at McGill revolves more around causal inference methods, especially frequentist estimation procedures. This difference in research expertise has allowed me to learn more about different statistical methods in translational medicine and familiarize myself with the direction in which research areas are currently heading.

The guidance and general advice provided by my supervisors, Dr. Erica Moodie and Dr. Rodolphe Thiébaut, have been invaluable. Firstly, from a collaborative and academic perspective, Dr. Moodie was always available to answer any statistics-related questions regarding in a succinct manner. While she is mostly known for her success in the DTR literature, what I admire most in her was her professionalism, work ethic and devotion to help her students. The same can be said about Dr. Thiébaut; underneath his bubbly personality is someone who deeply cares about his research and his students. He often had comments about clinical implications methodological assumptions in the design of the analysis and his frequent reminders made me realize that I do not think enough of the epidemiological aspect of public health research. In both of them, I particularly look up to their balance between their professionalism as a researcher and their "humanness" as an individual.

Above all, I am particularly grateful for the assistance and reassurance provided to me by the people within the statistical community when I had concerns to share. Whether I had questions regarding statistical methods or when I needed to talk about my desire to pursue a PhD, all the support and life advice that I have received from the community is what I appreciate most from this entire professional experience. Dr. Laura Villain, who defended her PhD thesis not long before my arrival in Bordeaux, was happy to share her knowledge regarding the data for this research project, which I am very thankful for. While I look forward to my upcoming PhD journey at the University of Toronto, it was the many conversations with my peers, professors and other members of my academic entourage that helped me figure out that helped me form my own decision. People with whom I sought advice regarding doctoral studies include Dr. Boris Hejblum, Dr. Alexandra Schmidt, Dr. Ilaria Montagni, Dr. Andrea Benedetti, Dr. Sahir Bhatnagar and Dr. David Stephens on top of my project supervisors. The many conversations with my peers Janie, Kristin, Chris, Thai-Son, Menglan, Peter, Armando, Paritosh, Kevin, Yi Lian, Guanbo, Steve, Iris, Sudip, Sneha and Arul amongst many others allowed me to finally straighten my thoughts and break the never-ending mental loop regarding my future plans. Although this emotional support and life advice are formally unrelated

to the academic aspect of my research project, they are what I will remember most from this entire experience.

References

[1] AUSTIN, P. C. Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. *Statistics in Medicine 37*, 11 (2018), 1874–1894.

[2] BELLMAN, R. A Markovian decision process. *Journal of Mathematics and Mechanics 6*, 5 (1957), 679–684.

[3] BELLMAN, R. Dynamic programming. *Science 153*, 3731 (1966), 34–37.

[4] CHAKRABORTY, B., LABER, E. B., AND ZHAO, Y. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics 69*, 3 (2013), 714–723.

[5] CHAKRABORTY, B., AND MOODIE, E. *Statistical methods for dynamic treatment regimes*. Springer: New York, USA, 2013.

[6] CYTHERIS. Cytheris announces initiation of ORVACS-sponsored phase II clinical study to attack viral reservoir of HIV patients. `http://www.natap.org/2010/newsUpdates/102210_04.htm`. Online; Accessed May 28, 2020.

[7] DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y., HEAGERTY, P. J., ZEGER, S., ET AL. *Analysis of longitudinal data*. Oxford University Press: New York, USA, 2002.

[8] DOUEK, D. C., ROEDERER, M., AND KOUP, R. A. Emerging concepts in the immunopathogenesis of AIDS. *Annual Review of Medicine 60* (2009), 471–484.

[9] EFRON, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[10] ENDO, G., MORIMOTO, J., MATSUBARA, T., NAKANISHI, J., AND CHENG, G. Learning CPG-based biped locomotion with a policy gradient method: application to a humanoid robot. *The International Journal of Robotics Research 27*, 2 (2008), 213–228.

[11] FLURY, B. K., AND RIEDWYL, H. Standard distance in univariate and multivariate analysis. *The American Statistician 40*, 3 (1986), 249–251.

[12] FRIEDEN, T. R. The future of public health. *New England Journal of Medicine 373*, 18 (2015), 1748–1754.

[13] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[14] GELBER, R. D., COLE, B. F., GELBER, S., AND GOLDHIRSCH, A. Comparing treatments using quality-adjusted survival: the Q-TWiST method. *The American Statistician 49*, 2 (1995), 161–169.

[15] GLASZIOU, P., SIMES, R., AND GELBER, R. Quality adjusted survival analysis. *Statistics in medicine 9*, 11 (1990), 1259–1276.

[16] GRABAR, S., LE MOING, V., GOUJARD, C., LEPORT, C., KAZATCHKINE, M. D., COSTAGLIOLA, D., AND WEISS, L. Clinical outcome of patients with HIV-1 infection according to immunologic and virologic response after 6 months of highly active antiretroviral therapy. *Annals of Internal Medicine 133*, 6 (2000), 401–410.

[17] HUBER, P. J., ET AL. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (1967), vol. 1, University of California Press, pp. 221–233.

[18] JARNE, A., COMMENGES, D., VILLAIN, L., PRAGUE, M., LÉVY, Y., THIÉBAUT, R., ET AL. Modeling CD4+ T cells dynamics in HIV-infected patients receiving repeated cycles of exogenous Interleukin 7. *The Annals of Applied Statistics 11*, 3 (2017), 1593–1616.

[19] KLEINBERG, J., AND TARDOS, E. *Algorithm design*. Pearson Education India, 2006.

[20] KOHL, N., AND STONE, P. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004* (2004), vol. 3, IEEE, pp. 2619–2624.

[21] KOSOROK, M. R., AND MOODIE, E. E. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, vol. 21. SIAM: Pennsylvania, USA, 2015.

[22] LAWSON, B. R., GONZALEZ-QUINTIAL, R., ELEFTHERIADIS, T., FARRAR, M. A., MILLER, S. D., SAUER, K., MCGAVERN, D. B., KONO, D. H., BACCALA, R., AND THEOFILOPOULOS, A. N. Interleukin-7 is required for CD4+ T cell activation and autoimmune neuroinflammation. *Clinical immunology 161*, 2 (2015), 260–269.

[23] LEVY, Y., SERETI, I., TAMBUSSI, G., ROUTY, J., LELIEVRE, J., DELFRAISSY, J., MOLINA, J., FISCHL, M., GOUJARD, C., RODRIGUEZ, B., ET AL. Effects of recombinant human Interleukin 7 on T-cell recovery and thymic output in HIV-infected patients receiving antiretroviral therapy: results of a phase I/IIa randomized, placebo-controlled, multicenter study. *Clinical Infectious Diseases 55*, 2 (2012), 291–300.

[24] LEWDEN, C., CHÊNE, G., MORLAT, P., RAFFI, F., DUPON, M., DELLAMONICA, P., PELLEGRIN, J.-L., KATLAMA, C., DABIS, F., LEPORT, C., ET AL. HIV-infected adults with a CD4 cell count greater than 500 cells/mm3 on long-term combination antiretroviral therapy reach same mortality rates as the general population. *JAIDS Journal of Acquired Immune Deficiency Syndromes 46*, 1 (2007), 72–77.

[25] LIU, N., LIU, Y., LOGAN, B., XU, Z., TANG, J., AND WANG, Y. Learning the dynamic treatment regimes from medical registry data through deep Q-network. *Scientific reports 9*, 1 (2019), 1–10.

[26] LOGEROT, S., RANCEZ, M., CHARMETEAU-DE MUYLDER, B., FIGUEIREDO-MORGADO, S., ROZLAN, S., TAMBUSSI, G., BEQ, S., COUËDEL-COURTEILLE, A., AND CHEYNIER, R. HIV Reservoir dynamics in HAART-treated poor immunological responder patients under IL-7 therapy. *AIDS 32*, 6 (2018), 715–720.

[27] MACKALL, C. L., FRY, T. J., AND GRESS, R. E. Harnessing the biology of IL-7 for therapeutic application. *Nature Reviews Immunology 11*, 5 (2011), 330–342.

[28] MOFENSON, L. M., BRADY, M. T., DANNER, S. P., DOMINGUEZ, K. L., HAZRA, R., HANDELSMAN, E., HAVENS, P., NESHEIM, S., READ, J. S., SERCHUCK, L., ET AL. Guidelines for the prevention and treatment of opportunistic infections among HIV-exposed and HIV-infected children: recommendations from CDC, the National Institutes of Health, the HIV Medicine Association of the Infectious Diseases Society of America, the Pediatric Infectious Diseases Society, and the American Academy of Pediatrics. *MMWR. Recommendations and Reports: Morbidity and Mortality Weekly Report. Recommendations and Reports/Centers for Disease Control 58*, RR-11 (2009), 1.

[29] MOODIE, E. A note on the variance of doubly-robust g-estimators. *Biometrika 96*, 4 (2009), 998–1004.

[30] MOODIE, E. E., CHAKRABORTY, B., AND KRAMER, M. S. Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics 40*, 4 (2012), 629–645.

[31] MOODIE, E. E., DEAN, N., AND SUN, Y. R. Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences 6*, 2 (2014), 223–243.

[32] NAHUM-SHANI, I., QIAN, M., ALMIRALL, D., PELHAM, W. E., GNAGY, B., FABIANO, G. A., WAXMONSKY, J. G., YU, J., AND MURPHY, S. A. Q-learning: A data analysis method for constructing adaptive interventions. *Psychological Methods 17*, 4 (2012), 478.

[33] OF THE COLLABORATION OF OBSERVATIONAL HIV EPIDEMIOLOGICAL RESEARCH IN EUROPE (COHERE) IN EUROCOORD, O. I. P. T. CD4 cell count and the risk of AIDS or death in HIV-infected adults on combination antiretroviral therapy with a suppressed viral load: a longitudinal cohort study from COHERE. *PLoS Medicine 9*, 3 (2012).

[34] PASIN, C., DUFOUR, F., VILLAIN, L., ZHANG, H., AND THIÉBAUT, R. Controlling IL-7 injections in HIV-infected patients. *Bulletin of Mathematical Biology 80*, 9 (2018), 2349–2377.

[35] PETERS, J., AND SCHAAL, S. Reinforcement learning of motor skills with policy gradients. *Neural Networks 21*, 4 (2008), 682–697.

[36] PETERSEN, M. L., DEEKS, S. G., AND VAN DER LAAN, M. J. Individualized treatment rules: generating candidate clinical trials. *Statistics in Medicine 26*, 25 (2007), 4578–4601.

[37] RAMASWAMI, R., BAYER, R., AND GALEA, S. Precision medicine from a public health perspective. *Annual Review of Public Health 39* (2018), 153–168.

[38] RICH, B., MOODIE, E. E., AND STEPHENS, D. A. Optimal individualized dosing strategies: a pharmacologic approach to developing dynamic treatment regimens for continuous-valued treatments. *Biometrical Journal 58*, 3 (2016), 502–517.

[39] ROBINS, J. M. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics* (2004), Springer, pp. 189–326.

[40] ROSENBAUM, P. R., AND RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 1 (1983), 41–55.

[41] ROSENDAHL, I., KIEBERT, G. M., CURRAN, D., COLE, B. F., WEEKS, J. C., DENIS, L. J., AND HALL, R. R. Quality-adjusted survival (Q-TWiST) analysis of EORTC trial 30853: Comparing goserelin acetate and flutamide with bilateral orchiectomy in patients with metastatic prostate cancer. *The Prostate 38*, 2 (1999), 100–109.

[42] SCHULZ, J., AND MOODIE, E. E. Doubly robust estimation of optimal dosing strategies. *Journal of the American Statistical Association*, just accepted (2020), 1–31.

[43] SIMONEAU, G., MOODIE, E. E., NIJJAR, J. S., PLATT, R. W., INVESTIGATORS, S. E. R. A. I. C., ET AL. Estimating optimal dynamic treatment regimes with survival outcomes. *Journal of the American Statistical Association* (2019), 1–9.

[44] STUART, E. A., LEE, B. K., AND LEACY, F. P. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology 66*, 8 (2013), S84–S90.

[45] SURH, C. D., AND SPRENT, J. Regulation of mature T cell homeostasis. *Seminars in Immunology 17*, 3 (2005), 183–191.

[46] SURH, C. D., AND SPRENT, J. Homeostasis of naive and memory T cells. *Immunity 29*, 6 (2008), 848–862.

[47] SUTTON, R. S., AND BARTO, A. G. *Reinforcement learning: an introduction*. MIT press, 2018.

[48] TAO, Y., WANG, L., AND ALMIRALL, D. Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *The Annals of Applied Statistics 12*, 3 (2018), 1914.

[49] THEODOROU, E., BUCHLI, J., AND SCHAAL, S. Reinforcement learning of motor skills in high dimensions: A path integral approach. In *2010 IEEE International Conference on Robotics and Automation* (2010), IEEE, pp. 2397–2403.

[50] THIEBAUT, R., DRYLEWICZ, J., PRAGUE, M., LACABARATZ, C., BEQ, S., JARNE, A., CROUGHS, T., SEKALY, R.-P., LEDERMAN, M. M., SERETI, I., ET AL. Quantifying and predicting the effect of exogenous Interleukin-7 on CD4+ T cells in HIV-1 infection. *PLoS Computational Biology 10*, 5 (2014).

[51] THIÉBAUT, R., JARNE, A., ROUTY, J.-P., SERETI, I., FISCHL, M., IVE, P., SPECK, R. F., D'OFFIZI, G., CASARI, S., COMMENGES, D., ET AL. Repeated cycles of recombinant human interleukin 7 in hiv-infected patients with low CD4 T-cell reconstitution on antiretroviral therapy: results of 2 phase II multicenter studies. *Clinical Infectious Diseases 62*, 9 (2016), 1178–1185.

[52] TIBSHIRANI, R. J., AND EFRON, B. An introduction to the bootstrap. *Monographs on statistics and applied probability 57* (1993), 1–436.

[53] VILLAIN, L., COMMENGES, D., PASIN, C., PRAGUE, M., AND THIÉBAUT, R. Adaptive protocols based on predictions from a mechanistic model of the effect of IL7 on CD4 counts. *Statistics in Medicine 38*, 2 (2019), 221–235.

[54] WALLACE, M. P., AND MOODIE, E. E. Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics 71*, 3 (2015), 636–644.

[55] YANG, D., AND DALTON, J. E. A unified approach to measuring the effect size between two groups using SAS. In *SAS global forum* (2012), vol. 335, pp. 1–6.

A. Appendix

List of Abbreviations

AIDS	Acquired immunodeficiency syndrome
BMI	Body mass index
CD4	CD4+ T cells, where CD4 is short for cluster of differentiation
DTR	Dynamic treatment regime
dWOLS	Dynamic weighted ordinary least squares
G-dWOLS	Generalized dynamic weighted ordinary least squares
HIV	Human immunodeficiency virus
IL-7	Interleukin 7
IPTW	Inverse probability treatment weights
ITR	Individualized treatment rule
LIR	Low immunological responders
PLWH	Person/people living with HIV
Q-TWiST	Quality-adjusted time without symptoms and toxicity
RL	Reinforcement learning
SD	Standard deviation
SMD	Standardized mean difference