Decision-Analytic Models for Treatment Optimization in the Presence of Patient Heterogeneity

Mutita Siriruchatanon

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Shan Liu, Chair

Horacio A. Duarte

Zelda B. Zabinsky

Program Authorized to Offer Degree:

Industrial & Systems Engineering

University of Washington

Abstract

Decision-Analytic Models for Treatment Optimization
in the Presence of Patient Heterogeneity

Mutita Siriruchatanon

Chair of the Supervisory Committee:
Associate Professor Shan Liu
Industrial & Systems Engineering

With the ever-increasing complexity in disease etiology, new therapeutics, healthcare service delivery, and clinical guidelines, selecting the appropriate course of treatment for an individual or population can become a great challenge for clinicians and healthcare providers. Applying suboptimal healthcare policies can set a damaging course for infectious disease control on the population level. On the individual level, disease can progress uniquely from patient-to-patient and ignoring a patient's preference may lead to treatment nonadherence or treatment rejection. In this thesis, we address the need for the development of decision-analytic methods for treatment selection that accounts for diversity in the patient population, uncertainty in patient treatment responses, and patients' preferences by studying the following problems:

(1) the HIV treatment policy selection in HIV-infected children in sub-Saharan Africa initiating treatment at age $\geq$ 3 years old in the presence of pre-treatment drug resistance;
(2) a personalized treatment selection problem for chronic depression where patient's respond uniquely to treatments whose response level is quantified by their unknown treatment effects;

(3) a personalized treatment selection problem with two competing objectives, health outcomes and treatment side effect burden, given the qualitative rankings of sequences of possible patient's treatment and responses

For the first problem, we develop and calibrate a microsimulation model of HIV disease progression and treatment. Using the model, we evaluate alternative antiretroviral treatment strategies using cost-effectiveness analyses. The second problem is formulated as a Markov Decision Process (MDP) where the treatment progression is parametrized by an individual's unknown treatment effects. We solved for the personalized treatment policies using two heuristic approaches: a model-based approach that can estimate an individual's treatment effect and a model-free approach using reinforcement learning. Taking into account an individual's preferences over two objectives, we formulate the last problem as an MDP as well. We developed two search algorithms, exhaustive and heuristic search, to estimate a patient's preference and provide an optimal treatment plan. This thesis contributes in developing three decision-analytic models to support decisions in testing, monitoring, and treatment selection for two significant healthcare problems, specifically, HIV and chronic depression, and treatment selection incorporating patient's preference. In addition, our work provides a step towards the design of personalized treatment strategy for patients with chronic diseases in various scenarios.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to my advisor, Dr. Shan Liu, who has always been supportive and guided me throughout my PhD journey. Over the past five years, she has provided constant encouragement, patience and invaluable advice that has built up to this dissertation. I am sincerely grateful for the opportunities she has given me that has provided me with the skill set I have today.

I am grateful to my supervisory committee members, Dr. Horacio Duarte, Prof. Zelda Zabinsky, and my GSR Prof. Shi Chen for their insightful feedback and guidance. I would like to especially thank Dr. Horacio Duarte for the opportunity to work on a topic that is substantial to my PhD thesis, along with his helpful advice, wisdom, and practical suggestions.

I also wish to thank the members of Prof. Liu's research group for their suggestions and support. Many thanks to all my friends from the department and the UW, especially Zahra and Daniel for always listening and helping. To my Thai friends, especially Jom, Mhee, and Pete, thank you for your mental support and delicious adventures. Lastly, I would like to thank my parents, Chairat and Sopana Siriruchatanon, for their love, understanding, and support throughout my entire life.

1. Introduction

1.1. Motivation

With the ever-increasing complexity in disease etiology, new therapeutics, healthcare service delivery, and clinical guideline, selecting the appropriate course of treatment for an individual or population can become a great challenge for clinicians and healthcare providers. Prescribing ineffective treatments, applying suboptimal healthcare policies, or ignoring patients' preferences can adversely affect the health outcomes of an individual or population. In the latter case, some policies can set a damaging course for infectious disease control, unnecessarily costing the lives of many [1].

With the emergence of big medical data in the electronic health records (EHR), medical decision makers can leverage these data and make evidence-based decisions to tailor treatment to an individual patient's disease diagnosis and prognosis, and preference. On the population level, policy makers can use decision-support tools to assist in designing cost-effective and life-saving healthcare interventions, which is especially important in resourced-limited settings. Likewise, new drugs that are introduced for the same disease may or may not have the same effectiveness in different populations. Thus, decision-analytic methods are needed to help design optimal treatment strategies to consider patient heterogeneity. On the individual level, disease can progress uniquely from patient-to-patient, where clinical guidelines should not be one-size-fits-all for everyone. Furthermore, incorporating a patient's preference may lead to better health outcomes as patients tend to have an increased satisfaction and adhere to their preferred treatments [2,3]. With advanced analytics, healthcare providers can select patient-specific treatment plan and possibly reduce medications' long-term side-effects. This thesis presents the development of decision-analytic models and solutions to guide testing, monitoring, and treatment decisions for two significant healthcare problems worldwide: HIV and chronic depression, and solutions to treatment selection while incorporating patient's preference.

1.2. Research objectives

The objective of this research is to develop decision-analytic methods that account for diversity in the patient population and uncertainty in patient treatment response with the aim of supporting treatment decisions for infectious and chronic diseases. To accomplish the research objective, we define the following tasks:

1) Provide policy guidance on selecting the optimal testing and treatment strategies for pediatric HIV patients in resourced-limited settings using a microsimulation model, cost-effectiveness analysis, and sensitivity analyses.
2) Develop personalized treatment strategies for chronic depression using model-free and model-based algorithms and evaluate their robustness towards different levels of patient heterogeneity.
3) Develop and evaluate search algorithms for estimating an individual's preference and providing an optimal policy in a personalized treatment selection problem with two competing objectives given the individual's trajectory rankings.

The three tasks combined provide a broad view of possible implementations for medical decision-making tools from the population level down to the personalized level.

1.3. Organization of this thesis

This thesis is organized as follows: In Chapter 2, we focus on the implementation of HIV treatment policies using a microsimulation model and outcome analyses including cost-effectiveness analyses, sensitivity analyses, and scenario analyses. The detail of the development and calibration of a microsimulation model for HIV-infected children in Kenya is included in Appendix A. In Chapter 3, we consider a personalized treatment selection problem for chronic depression, which is formulated as a Markov Decision Process, and apply model-free and model-based algorithms to our problem. Lastly, we present a personalized treatment selection problem with two competing objectives in the presence of an unknown individual's preference and introduce two search algorithms for estimating the preference and providing an optimal policy in Chapter 4.

Chapter 2 provides an overview of a microsimulation model of disease progression and treatment in HIV-infected children in sub-Saharan Africa where there is high prevalence of pretreatment drug resistance. Using this model, we evaluated several antiretroviral therapy (ART) strategies for children initiating ART from the age of 3 years in settings where the recently approved Dolutegravir (DTG)-based ART is and is not available as the preferred first-line regimen for children. In the settings without DTG, three strategies were included in evaluation: 1) status quo with empiric NNRTI-based first-line ART (*status quo*); 2) using PDR testing with consensus sequencing in guiding ART initiation (*PDR testing*); 3) improving rate of switching to second-line ART (*improved switching*). In the settings with DTG, two strategies were included in the evaluation: 4) introducing DTG as a first-line ART (*DTG status quo*); and 5) a combination of DTG as a first-line ART and improving rate of switching to second-line ART (*DTG improved switching*). To evaluate the strategies in the two settings, we compared health outcomes, costs, and cost-effectiveness of these strategies over a 10-year time horizon. Results showed that *improved switching* has an incremental cost-effectiveness ratio (ICER) of US$579/LY compared to *status quo* and *PDR testing* is dominated by *improved switching*. In the settings with DTG, *DTG improved switching* has an ICER of US$591/LY compared to *DTG status quo*.

Chapter 3 presents a personalized treatment selection problem for chronic depression where patient heterogeneity lies in their treatment response types, i.e. patients are assumed to respond more towards one treatment than others. We formulated a treatment selection problem as a Markov Decision Process and developed two heuristic algorithms with the objective of maximizing patients' health outcomes. The two heuristic algorithms are one-step look-ahead with individual treatment effect estimation (a model-based approach) and Fitted Q-Iteration (a model-free approach). We evaluated the proposed algorithms on two experimental setups. The first setup assumes patients responding to the same treatment has the same treatment effect values. In the second setup, we increased patient heterogeneity in treatment response by varying the treatment effect values among those who respond to the same treatment. Our simulation results showed that the model-based algorithm is more accurate in providing the correct treatment and can capture heterogeneity in both setups compared to the model-free algorithm. Comparing the performance of the model-free algorithm between the two setups, we observed an improvement in providing

more of the correct treatments with increasing patient heterogeneity (setup two), while the model-based algorithm performed better in low patient heterogeneity (setup one).

Lastly, Chapter 4 proposes two search algorithms, exhaustive and heuristic search, for estimating an individual's preference and providing optimal policies in a personalized treatment selection problem with two competing objectives, given the trajectory rankings from individual and access to an expert's trajectories. The problem is formulated as a Markov decision process where the true reward function is a linear combination of preferences between two competing objectives: health outcomes and side effects. The exhaustive search considers all possible preference weights in the search space and generates their corresponding optimal policies, while the heuristic search samples from the search space with the objective of finding preference weights that can and cannot match the provided trajectory rankings. Both algorithms eliminate any infeasible weights and iteratively reduce the search space. To evaluate our algorithms' performance, we conduct numerical experiments on combinations of an individual's preference and initially incorrect preference weights at baseline and performed one-way sensitivity analysis. Our results show that both algorithms successfully converge to the true preference weight with small errors. The heuristic search performs better in all metrics including the accuracy of preference weight estimation, optimality gap, number of iterations to terminate the algorithm, and runtimes.

2. Cost-effectiveness analysis of pediatric ART policy options in the presence of pretreatment NNRTI drug resistance in sub-Saharan Africa

2.1. Introduction

As of 2020, there are approximately 1.7 million children living with HIV globally with 89% living in sub-Saharan Africa alone. From 2010 to 2018, there has been a 64% decline and only a 36% decline in new HIV infections in children from Eastern and Southern Africa and from West and Central Africa, respectively. In 2020, only 77% and 73% of children living with HIV were on treatment in Eastern and Southern Africa and from West and Central Africa [4], which is still severely far from achieving the pediatric UNAIDS 90-90-90 goal.

One of the issues possibly contributing to the low viral suppression in children is pretreatment drug resistance (PDR) as the PDR prevalence in sub-Saharan African children has been increasing [5]. A meta-analysis study on PDR in sub-Saharan African children [5] reported that the pooled PDR prevalence among children exposed to the prevention of mother-to-child transmission (PMTCT) and PMTCT-unexposed were 43% and 13%, respectively. The common drug resistance is towards non-nucleoside reverse transcriptase inhibitor (NNRTI)-based treatment, which is contained in drugs for PMTCT [5], and is one of the antiretroviral therapy (ART) regimen choices for children [6]. Since the PMTCT coverage in Eastern and Southern Africa is 95% and in West and Central Africa is 56% as of 2020 [4] and PDR is associated with an increase in the risk of virologic failure [7,8], PDR is likely one of the major factors of the high virologic failure outcomes in sub-Saharan African children.

In 2013, the WHO began recommending protease inhibitor (PI)-based ART as the first-line regimen for children < 3 years old, with NNRTI-based ART as the recommended first-line regimen for children ≥ 3 years old [6]. When treatment failure occurs, children who initiate ART with NNRTI-based ART can switch to the second-line ART, PI-based ART [6], which is currently the last available ART regimen for HIV-infected children in sub-Saharan Africa. However, more than 50% of HIV-infected children in sub-Saharan Africa do not initiate ART by the age of 3 years old [9,10], thus, using NNRTI- based ART leaves children in this age group with PDR at increased risk of virologic failure.

Dolutegravir (DTG) was recently approved and is now recommended for use in children as first-line ART [11]. Several studies have proved that DTG-based ART has better virologic efficacy and has a higher-resistance barrier than NNRTI-based ART [12–14], and therefore, will help address the challenges associated with PDR. However, there may be barriers to rapid roll-out of dolutegravir-based ART for children, which could result from an unprepared supply chain for the transition to DTG-based first-line ART or the accessibility to low-cost generic DTG [15]. In adults, the low-cost generic DTG is expected to be $75 per person per year [16], however, the price of the dispersible-tablet formulation of DTG for use in infants and children has not been announced as of the time of writing.

To address the uncertainty in DTG rollout and the challenges associated with PDR, testing for PDR is one potential strategy that could improve health outcomes for children ≥ 3 years old. With PDR testing, children with NNRTI resistance would gain benefits from initiating ART with PI-

based ART. While various studies suggest that testing for PDR to NNRTI-based ART is not cost-effective for adults in resource-limited settings [17,18], it is important to evaluate the cost-effectiveness of this intervention in children, as meta-analyses suggest that rates of virologic failure associated with PDR are higher among children compared to adults in resourced-limited settings [7,8]. In sub-Saharan Africa, a relatively low rate of switching to second-line ART has been observed among children [19,20]. The study prior to the availability of DTG on adults in the presence of PDR has shown that increasing the rate of switching as an intervention is cost-effective. Increasing the rate of switching to second-line ART when virologic failure is diagnosed is another strategy that could also improve health outcomes and is likely to be cost-effective in the settings where DTG is not available.

Therefore, we aim to evaluate the effectiveness and cost-effectiveness of ART strategies for HIV-infected children initiating ART at 3 years and older in sub-Saharan Africa where the prevalence of PDR is high. We developed and calibrated a microsimulation model of disease progression and treatment in HIV-infected children in sub-Saharan Africa. Using this model, we implemented five alternative ART strategies for children initiating ART from the age of 3 years and compared health outcomes, costs, and cost-effectiveness of these strategies over a 10-year time horizon, in settings with and without DTG availability.

2.2. Methods

2.2.1. Model

We developed a microsimulation model that simulates the disease progression and treatment of HIV-infected children in the presence of pretreatment NNRTI drug resistance (Appendix A-I). The model follows the disease progression of chronic HIV-infected children from the age of 0 to 13 years where each child is tracked on CD4%/CD4 cell count, viral load, ART regimens, ART duration, ART failure, opportunistic infections (OI) history, and drug resistance status. The main immunologic measure when the population is under age 5 is CD4% while absolute CD4 cell count is used thereafter. A detailed conversion of CD4% to CD4 cell count at age 5 can be found in Appendix A-II. Without treatment, the CD4 level declines each month. The rate of CD4 decline in children < 5 is constant, while the rate in children aged ≥ 5 depends on the viral load level. Lower CD4 is associated with a higher risk of death. In addition, CD4 level determines the risk of OI in children aged ≥ 5. The parameters relevant to the disease progression for children aged 0–4 years and for children aged 5–13 years are presented in Table 2-1 and Table 2-2, respectively.

In this model, ART is initiated when children aged of 3 [9,10] are diagnosed with HIV infection regardless of CD4 level [21]. For those on treatment, virologic failure can occur with a probability depending on ART, the duration on ART, and drug resistance status where the presence of drug resistance is associated with a higher probability of failure on NNRTI-based ART [7,8]. With effective ART, CD4 level increases according to the ART regimen and viral load is suppressed, which in turn decreases the risks of OI and death. The parameter related to effective ART is specified in Appendix A Table 2. Additional detail on the model is in Appendix A-I.

To simulate a realistic sub-Saharan HIV-infected pediatric population, the model was calibrated against observational targets including UNAIDS survival outcomes of untreated HIV infected

children from age 0 to 5 years [22,23], the statistics of CD4% and CD4 cell count of HIV-infected children not on ART at age 5 (Appendix A Table 5), and the P1060[1] observed mortality rate and opportunistic rate among HIV-infected children on ART [24]. The parameters included in our calibration are specified in Appendix A Table 4 and Appendix A Table 5. All included parameters are calibrated simultaneously by considering the population of children on treatment and not on treatment, which results in the calibrated parameters presented in Table 2-1 and Table 2-2. The calibration process is described as follows:

1. Generate a unique 20,000 parameter sets where each set consists of 32 parameters by randomly sampling from the predefined distribution (Uniform and Normal distributions) for each parameter (Appendix A Table 4 and Appendix A Table 5).
2. Run each parameter set through the model by simulating a population of 200,000 children and follow them from birth to 5 years old. The population consists of the following:
    a. The "No ART" cohort: 100,000 children who are not on treatment and are used in generating UNAIDS survival outcomes and CD4 statistics at age 5
    b. The "ART" cohort: 100,000 children who initiate ART at age of 1 year are used in generating the P1060 mortality and OI rate. Within the ART cohort, children were randomized to two arms, nevirapine (NVP) and lopinavir (LPV), where each arm has 50,000 children. The two arms have different treatment options as follows:
        i. NVP arm: NNRI based first-line ART followed by PI-based second-line ART
        ii. LPV arm: PI based first-line ART followed by NNRTI-based second-line ART

        Both arms assumed 100% switch to second-line ART if virologic failure occurs. The probabilities of viral load suppression over 12 months by ART regimen for each arm are specified in Appendix A Table 7.
3. For each parameter set, calculate the goodness-of-fit, mean squared error (MSE), for each target outcome where MSE is scaled by the target minimum and maximum bounds (specified in Appendix A Table 6). Then, calculate the mean MSE over all targets for each set.
4. Eliminate any parameter sets that provide outcomes that are out of minimum and maximum bounds.
5. Rank parameter sets using the mean MSE and obtain the best-fitted 50 parameter sets

We select the set with the lowest mean MSE as our model parameters. Full detail on calibration targets, parameter inputs, and calibration results can be found in Appendix A-II.

[1] P1060 study is a clinical trial for comparing antiviral responses to NNRTI-based and PI-based therapy in HIV-infected infants who have and have not received single dose nevirapine.

Table 2-1. Parameters relevant to HIV disease progression in children aged 0–4 years old

Parameters	**Parameter Value**			**Source**
Initial mean CD4% at age 0 (SD)	44.2% (10.0%)			Calibrated
Monthly rate of CD4 decline by age (%)				
≤ 3 months	6.2%			Calibrated
4 months – 5 years	0.5%			Calibrated
Monthly probability of clinical event by age				
< 6 months	3.3%[a]			[22]
6–59 months	4.6%[a]			[22]
Monthly probability of death with no history of clinical event by age	CD4%			
	< 15%	15-24%	> 25%	
0–6 months	6.3%	7.1%	6.9%	Calibrated
7–12 months	4.6%	5.2%	5.0%	Calibrated
13–24 months	1.5%	1.7%	1.6%	Calibrated
25–36 months	1.6%	1.8%	1.8%	Calibrated
37–48 months	0.4%	0.5%	0.5%	Calibrated
49–60 months	0.1%	0.1%	0.1%	Calibrated
Probability of death within 30 days of clinical events	11.1%			Calibrated
Monthly probability of death with history of clinical event (> 30 days post-event) by age	CD4%			
	< 15%	15-24%	> 25%	
0–6 months	22.3%	7.9%	3.9%	Calibrated
7–12 months	11.7%	4.0%	1.9%	Calibrated
13–24 months	6.2%	2.1%	1.0%	Calibrated
25–36 months	10.4%	3.5%	1.7%	Calibrated
37–48 months	1.2%	0.4%	0.2%	Calibrated
49–60 months	4.4%	1.5%	0.7%	Calibrated

a) Clinical events in our model include WHO Stage 3, Stage 4, and Tuberculosis events for HIV-related disease in children as defined by WHO [31]. The monthly probability of clinical events are the average values of the three aforementioned events.

Table 2-2. Parameters relevant to HIV disease progression in children aged 5–13 years old

Parameters	Parameter Value			Source
Initial mean CD4 cell count at age 5 (SD)	547.2 (311.6)			Calibrated
Monthly rate of CD4 cell count decline for age ≥ 5 by log_{10} viral load level (cells/μL)				
0 - 2.6	0			[17]
2.6 - 3.7	4.40			[17]
3.7 - 4.5	5.50			[17]
> 4.5	6.60			[17]
CD4 drop during acute OI for age ≥ 5 (cells/μL)	58.54			[17]
Monthly probability of clinical event by age				
5-9 years	2.3%			Appendix A-I
10-14 years	1.0%			Appendix A-I
Monthly probability of death with no history of clinical event by age	CD4 cell count			
	< 200	200-499	≥ 500	
5-9 years	0.2%	0.3%	0.3%	Appendix A-I
10-14 years	0.1%	0.1%	0.1%	Appendix A-I
Probability of death within 30 days of clinical events by age				
5-9 years	3.7%			Appendix A-I
10-14 years	2.1%			Appendix A-I
Monthly probability of death with history of clinical event (> 30 days post-event) by age	CD4 cell count			
	< 200	200-499	≥ 500	
5-9 years	1.8%	0.6%	0.3%	Appendix A-I
10-14 years	1.0%	0.3%	0.2%	Appendix A-I

2.2.2. Treatment strategies

We evaluated the cost-effectiveness of alternative ART strategies for HIV-infected children initiating ART ≥ 3 years old in two settings where dolutegravir (DTG)-based ART is and is not available. In the settings without DTG availability, we evaluated three alternative strategies: 1) the standard of care with empiric NNRTI-based first-line ART (*status quo*), 2) PDR testing to guide choice of initial ART regimen (*PDR testing*); 3) increasing rate of switching to protease inhibitor (PI)-based second-line ART (*improved switching*). In the settings with DTG availability, the empiric NNRTI-based first-line ART is replaced with DTG-based ART, which is associated with a lower probability of virologic failure compared to NNRTI-based first-line ART. As the rates of PDR to DTG is currently low, we do not include a strategy with PDR testing in our evaluation. Therefore, we considered two alternative strategies: 4) DTG-based ART as the empiric first-line regimen (*DTG status quo*); and 5) DTG-based ART as the empiric first-line regimen combined with increasing rate of switching to second-line ART (*DTG improved switching*).

In all five strategies, we assumed only two lines of ART are available, namely, NNRTI-based/DTG-based first-line ART and PI-based second-line ART. In the *status quo* and *improved*

switching strategies, children initiate ART with the empiric NNRTI regardless of their PDR status and switch to PI-based ART if virologic failure occurs. The rate of switch to the PI-based ART in the *status quo* and *improved switching* strategies are 40% and 80%, respectively. In the *PDR testing* strategy, children are tested with PDR testing before ART initiation. For those whose results are positive, they start the treatment with PI-based ART. Otherwise, they follow the same treatment strategy as in the *status quo* strategy. The *DTG status quo* and *DTG improved switching* strategies are identical to the *status quo* and *improved switching* strategies aside for the use of DTG-based ART as the empiric first-line regimen.

For all strategies, children always remain on treatment after ART initiation unless they are lost to follow-up (LTFU). We assumed that no LTFU after children are on treatment for a duration of 5 years due to lack of data. The study on several HIV-infected adult cohorts reported that patients gain more benefit by continuing on a failed treatment than completely discontinuing the treatment [25]. Therefore, we assume that individuals will continue on PI-based ART, a second-line ART, regardless of virologic failure. All parameters relevant to the alternative ART strategies are specified in Table 2-3 and Table 2-4.

2.2.3. Sensitivity and scenario analyses

To study the effect of uncertainty in the model parameters on the outcomes, we performed one-way sensitivity analyses on important parameters including PDR prevalence, the probability of virologic failure with PDR on NNRTI-based ART, the probability of virologic failure on PI-based ART, the rate of switching to second-line ART for strategies with improved switching, and various costs of HIV care. The ranges for sensitivity analyses are shown in Table 2-3 and Table 2-4. In addition, we conducted the scenario analyses focusing on a cost incurred from improving the rate of switching to second-line ART. Specifically, we consider a combination of different costs per child with diagnosed virologic failure on first-line ART and different levels of the rate of switching to second-line ART and evaluated the changes in the cost-effectiveness of the *improved switching* and *DTG improved switching* strategies.

Table 2-3. Parameters relevant to ART strategies for base-case analysis and sensitivity analysis

Parameters	Parameter Value	Range for sensitivity analyses	Source
PDR prevalence	18%	5–30%	[5]
Probability of virologic failure on initial ART (over 12 months)			
No PDR on NNRTI-based ART	19.2%[a]	16.8–24.6%	Appendix A-II
PDR on PI-based ART	19.2%		Appendix A-II
PDR on NNRTI-based ART	64.1%[a]	39.5-75.2%	Appendix A-II
DTG as first-line ART	9.10%[b]		Appendix A-II
Probability of virologic failure on second-line ART (over 24 months)			
PI-based ART after NNRTI-based ART	16.4%	13.9% - 19.4%	[26]
PI-based ART after DTG-based ART	16.4%	13.9% - 40.0%	Assumed
Probability of switching to second-line ART by 1 year after virologic failure is diagnosed by strategy			
Status quo/PDR testing/DTG status quo	40%		Assumed
improved switching/DTG improved switching	80%	60.0%–90%	
Probability of lost to follow-up (over 5 years)	15%		[27]

a) In the base-case analysis, the failure rate of NNRTI-based ART in children with PDR was calculated from the base-case values of PDR prevalence, the rate of viral suppression in children on first-line ART after 12 months [28], and the ratio of the odds for virologic failure of children with PDR on NNRTI-based ART compared to those without PDR on NNRTI-based ART (Odds ratio = 7.5). The ranges for sensitivity analyses were based on the odds ratios of 2 and 15 [7]. b) In the base-case analysis, the failure rate of DTG-based ART in children was estimated from the ratio of the odds for virologic failure of HIV-infected adults without PDR on NNRTI-based ART compared to those on DTG-based ART (Odds ratio = 2.37) due to the lack of available information. The ranges for sensitivity analyses were based on 0.5 times and 3 times of the adult odds ratio.

Table 2-4. Parameters relevant to costs for base-case analysis and sensitivity analysis

Parameters	Parameter Value	Range for sensitivity analyses	Source
Inpatient and outpatient relevant costs			
Inpatient costs per visit	$96	$15–$400	[29–33]
Outpatient costs per visit	$32	$10–$80	[29–33]
Number of visits for patients with OI & on ART	3		[17]
Number of visits for patients with OI & not on ART	7		[17]
Cost of testing per test			
CD4 testing	$12	$6–$24	[17]
Viral load testing	$54	$10–$80	[17]
Resistance testing	$125	$30–$250	[17]
Cost of ART per person per year			
NNRTI/DTG[a]	$123		[4]
PI	$290	$123–$400	[4]

a. The cost of DTG per person per year for children is assumed to be the same as NNRTI due to the lack of available information. The cost of DTG in adults is less costly than that of NNRTI [4], our assumption would be a worst-case scenario assuming that DTG is less costly in children as well.

2.3. Results

2.3.1. Calibration

As previously mentioned, our model was calibrated against several observational targets and the best-fit parameter set was selected using MSE as a metric for goodness-of-fit. Using the selected parameter set, our calibrated model produces outputs matching all calibration targets. The model outputs for UNAIDS survival outcomes, P1060 mortality rate and the statistics of CD4% and CD4 cell count of HIV-infected children at age 5 deviates from the targets with a MSE over all targets of less than 0.01 (Appendix A-II Section 1). Figure 2-1 shows that the model-generated survival curve is consistent with UNAIDS survival curve.

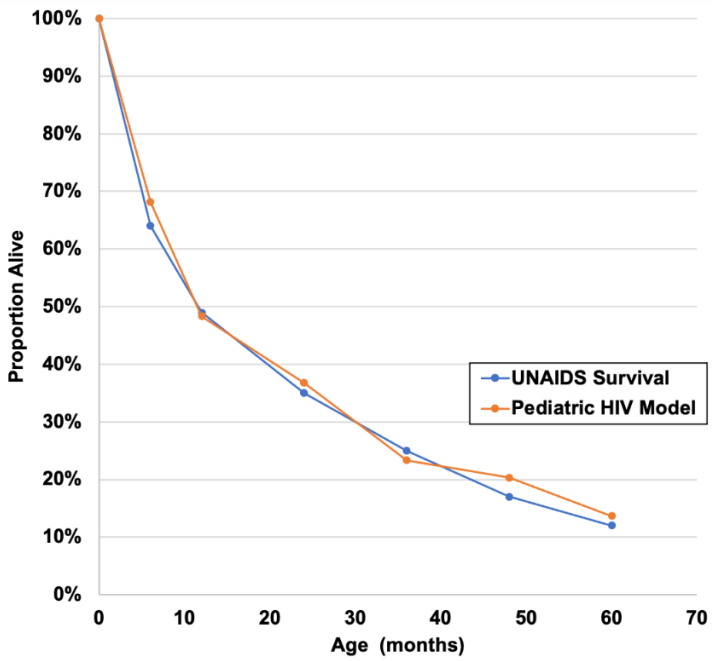


Figure 2-1. Model outputs for survival outcomes

The observational survival outcomes from UNAIDS (blue lines) is on the untreated HIV-infected children cohort from the age of 0 to 60 months [22,23]. The children cohort include more than 1,300 children from eight sub-Saharan Africa. The model-generated survival outcomes (orange line) have a small deviation from the UNAIDS observed survival curves with the MSE over all data points less than 0.02.

2.3.2. Health outcomes and costs

When DTG-based first-line ART is not available, the *improved switch* strategy provides the highest improvement in the proportion of children with suppressed viral load of at least 3% (66.2% vs. 63.7%) compared to the *status quo*. Comparing the *PDR testing* strategy to the *status quo*, *PDR testing* strategy shows an improvement of 2% approximately (65.0% vs. 63.7%). Similar results are shown in the proportion of children alive, where the *improved switch* strategy and the *PDR testing* strategy provided improvements of 2.8% (71.2% vs. 69.3%) and 2.5% (71.0% vs. 69.3%), compared to the status quo, respectively. In the settings without DTG availability, *DTG improved switching* strategy provided a higher proportion of children with suppressed viral load (VL) and proportion of children alive compared to the *DTG status quo*, which resulted in improvements of 1.3% (68.3% vs. 67.4%) and 0.9% (72.1% vs. 71.4%), respectively.

Due to a higher proportion of children on PI-based ART, there is an increase in the use of ART, specifically PI-based ART, in *improved switch* and *PDR testing* strategies, compared to *status quo* (Table 2-5). As PI-based ART is more costly than NNRTI-based ART, the costs of ART increased by 81% and 51% relative to the total costs in *improved switch* and *PDR testing* strategies, compared to *status quo* (Figure 2-2). By being on effective ART, the proportion of children alive is slightly higher in *improved switch* and *PDR testing*, which resulted in a relative increase in costs of viral load testing and outpatient care by 2.2% and 0.9%, respectively, compared to the *status*

quo. As resistance testing was applied to every child in the population, it incurred a cost increase relative to the total costs of 46%, compared to the *status quo*, leading to *PDR testing* being the most costly strategy. Overall, *improved switch* and *PDR testing* strategies have higher total costs compared to *status quo* by 4% and 14%, respectively.

Comparing between *status quo* and *DTG status quo*, *DTG status quo* has higher use of ART (Table 2-5), however, children were less likely to have a virologic failure in their first-line ART leading to lower use of PI-based ART. Being on effective ART is associated with lower risk of clinical events and risk of mortality. Therefore, *DTG status quo* has higher proportion of children alive and children with suppressed VL, which incurred an increase in costs of viral load testing and outpatient care by 1.6% and a decrease in inpatient care by 11% (Figure 2-2). Therefore, *DTG status quo* has lower total costs compared to *status quo* by 5%.

Table 2-5. Health and ART outcomes of the five alternative ART strategies

Health and ART outcomes	Status quo	Improved switching	PDR testing	DTG Status quo	DTG Improved switching
Proportion of children with suppressed viral load at 5 years after ART initiation[a]	63.7%	66.2%	65.0%	67.4%	68.3%
Proportion of children alive at 5 years after ART initiation[b]	69.3%	71.2%	71.0%	71.4%	72.1%
Proportion of children on PI-based ART[c]	17%	21%	26%	5%	7%
Person-months of ART use[d] (per person)	82.3	84.1	83.5	84.4	85.0
Person-months of PI-based ART use[d] (per person)	13.7	17.4	22.5	4.5	5.7

a) The numerator is the number of children with suppressed viral load at 5 years after ART initiation. The denominator includes all children who initiated ART at age 3 years, which by 5 years after ART initiation includes children with viral suppression, children who have been lost to follow-up, and children who have died. b) The numerator is the number of children alive at 5 years after ART initiation. The denominator includes all children who initiated ART at age 3 years. c) Average over 10-year time period. d) Total per person over 10-year time period.

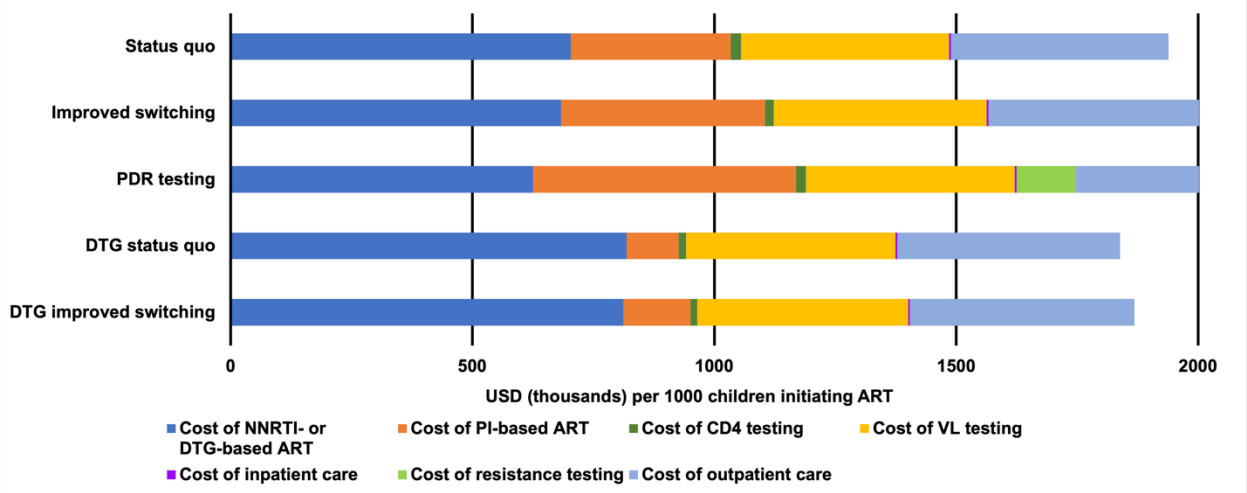


Figure 2-2. Cost breakdown: Total undiscounted costs for each strategy reported in per 1000 children initiating ART over a 10-year time period.

2.3.3. Cost-effectiveness

In settings without DTG availability (strategies in blue of Figure 2-3), the *PDR testing* strategy is dominated by the *improved switching* strategy. Compared to the *status quo* strategy, the *improved switching* strategy provides an additional 131 discounted LYs at an additional discounted cost of $75,592 per 1,000 children initiating ART as shown in Table 2-4 (an incremental cost-effectiveness ratio (ICER) of $579/LY gained). When DTG-based ART is available (strategies in red of Figure 2-3), the *DTG improved switching* strategy provides an additional 44 discounted LYs at an additional discounted cost of $25,746 per 1,000 children initiating ART (an ICER of $591/LY gained), compared to the *DTG status quo* strategy. Comparing between strategies with and without DTG, the strategies with DTG-based first-line ART provided higher benefit at lower costs than the strategies with NNRTI-based first-line ART. If we considered a traditional cost-effectiveness threshold, the 2020 sub-Saharan Africa gross domestic per capita (GDP) of US$850 [34], the *improved switching* strategy and *DTG improved switching* strategy are considered cost-effective.

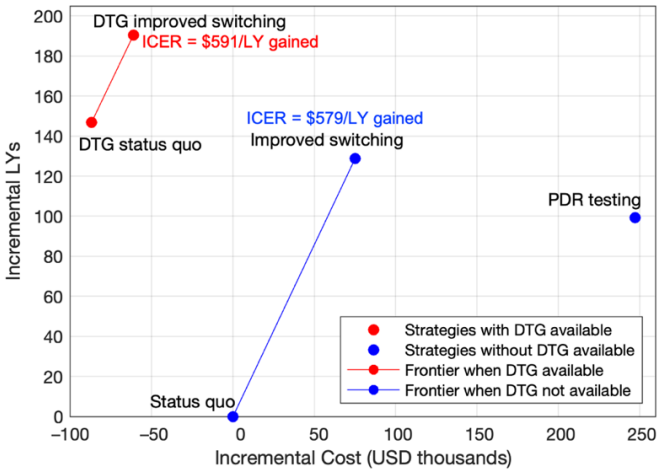


Figure 2-3. Incremental cost and benefits of the alternative ART strategies compared to the status quo. Incremental cost and benefits are per 1000 children initiating ART over a 10-year time period.

Table 2-6. Costs, LYs, and incremental cost-effectiveness of each strategy

	Undiscounted Cost ($US)	Undiscounted LYs	
Status quo	1,938,996	7,203	
Improved switching	2,025,987	7,358	
PDR testing	2,203,694	7,318	
DTG status quo	1,838,619	7,378	
DTG Improved switching	1,868,298	7,430	
	Discounted Cost ($US)	Discounted LYs	ICER
Status quo	1,697,253	6,301	
Improved switching	1,772,844	6,432	579
PDR testing	1,944,011	6,399	
DTG status quo	1,610,327	6,448	
DTG Improved switching	1,636,073	6,491	591

2.3.4. One-way sensitivity analyses

In both settings, the *status quo* and strategies with improved switching are the only two strategies on the frontier, therefore, we focused on analyzing and understanding how their ICERs were affected by each key parameter. The results from the one-way sensitivity analyses indicate that the cost of PI-based ART, the cost of outpatient care, and the cost of viral load testing are major factors driving our model in settings with and without DTG availability (Figure 2-4). Increasing the cost of PI-based ART, outpatient care, viral load testing, and the virologic failure rate on PI-based ART increased the ICER associated with *improved switching* compared to *status quo*, whereas, increasing cost of CD4 testing, PDR prevalence, cost of inpatient care, the rate of switch, and the probability of virologic failure on NNRTI-based ART with PDR decreased the ICER values (Figure 2-4A). In settings with DTG availability, we observed similar results as well (Figure 2-4B).

Due to the higher proportion of children on PI-based ART in strategies with improved switching compared to their status quos, increasing the cost of PI-based ART increased the ICER. As the incremental cost increases while the incremental benefit remains the same, it is not surprising that ICER increased as the cost of viral load testing or outpatient care increased. In our model, CD4 cell count is tested when virologic failure is detected. As strategies with improved switching have higher proportions of children with suppressed VL, the total cost of CD4 testing is lower than that of status quo as shown in the base-case result (Figure 2-2). When the unit cost of CD4 testing increases, the total cost of CD4 testing in status quo increases at a higher rate compared to strategies with improved switching. Consequently, the incremental cost decreases while the incremental benefit remains the same leading to a decrease in ICER. In strategies with improved switching, children experienced less incidents of OI compared to their status quo, which in turn required less inpatient care. By avoiding the cost of inpatient care, strategies with improved switching provided more benefit at a lesser cost, and consequently, increasing the cost of inpatient care decreased the ICER.

It is as expected that increasing the virologic failure rate of PI-based ART increased the ICER due to more children on PI-based ART losing benefit from staying on ineffective ART. Increasing the

rate of switching to a second-line ART in strategies with improved switching decreased the ICERs because the rate of benefit gained in *improved switching/DTG improved switching* is higher than the rate of increased cost.

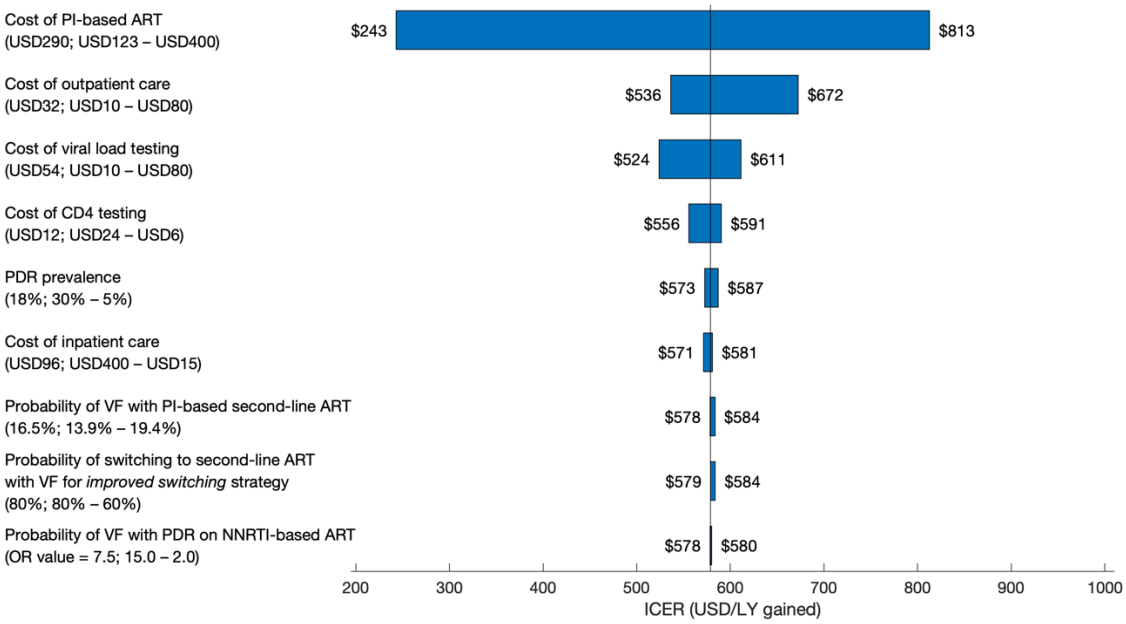
In settings without DTG availability, an increase in PDR prevalence leads to a larger proportion of children having PDR to NNRTI-based ART, which leads to a larger proportion of children failing on the first-line ART in both *improved switching* strategy and *status quo*. As virologic failure rate of NNRTI-based ART for children with PDR increases, those with PDR are more likely to fail on the first-line ART, which resulted in more children failing on the first-line ART on both strategies. However, a proportion of children who switched to a second-line ART out of the children eligible to switch (those who failed on the first-line ART) is smaller in the *status quo* compared to the *improved switching* strategy, therefore the ICER increased as the PDR prevalence or the virologic failure rate of NNRTI-based ART increased.

2.3.5. Scenario Analyses

In our base-case analyses, we assumed that there is no cost incurred from improving the rate of switching to second-line ART in strategies with improved switching. However, we believe that there is such a cost in practice. Therefore, we introduced "cost per child diagnosed with virologic failure (VF)" to *improved switching* and *DTG improved switching* strategy: a cost associated with a person who is diagnosed with virologic failure on the first-line ART and, therefore, is eligible to switch to a second-line ART. We explored the changes in the ICER values by simultaneously varying the improved rate of switch and the cost per child diagnosed with VF.

According to Figure 2-5 A, the ICER of *improved switching* compared to *status quo* decreased as the cost per child diagnosed with VF decreased and the improved rate of switch increased. Similar results are observed in the changes of the ICER of *DTG improved switching* compared to *DTG status quo* as well (Figure 2-5 B). If we consider the worst-case scenario where cost per child diagnosed with VF is at $120, the ICERs of *improved switching* compared to *status quo* are $1186/LY and $764/LY when the improved rates of switching are 50% and 90%, respectively. Similarly, the ICERs of *DTG improved switching* compared to *DTG status quo* are $1127/LY at 50% improved rates of switching and $758/LY at 50% improved rates of switching. It is worth mentioning the ICERs of strategies with improved rate of switching compared to their status quos are lower than $US850/LY at all costs per eligible person when the improved rate of switch is 60% and the cost per child diagnosed with VF is no more than $90. In addition, when the probability of virologic failure on PI-based second-line ART increases from 16.4% to 40% over 24 months, the ICER of *DTG improved switching* compared to *DTG status* shows a slight increase (Appendix A Figure 3 in Appendix A-III). These results imply that improving the rates of switching to second-line ART is a promising strategy in a resource-limited setting.

A.



B.

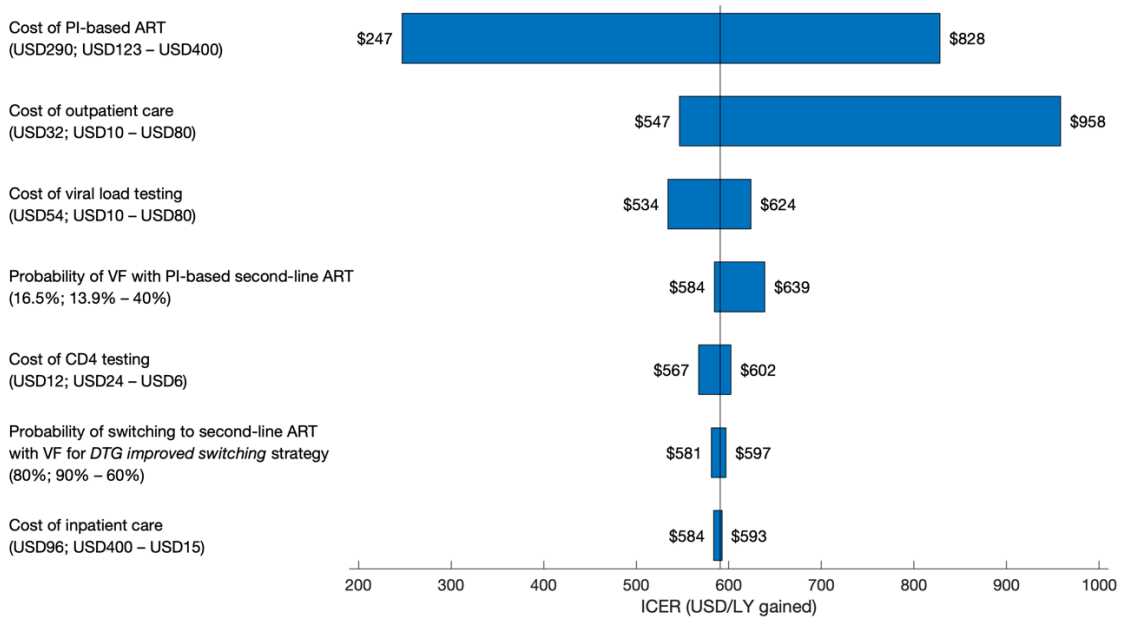
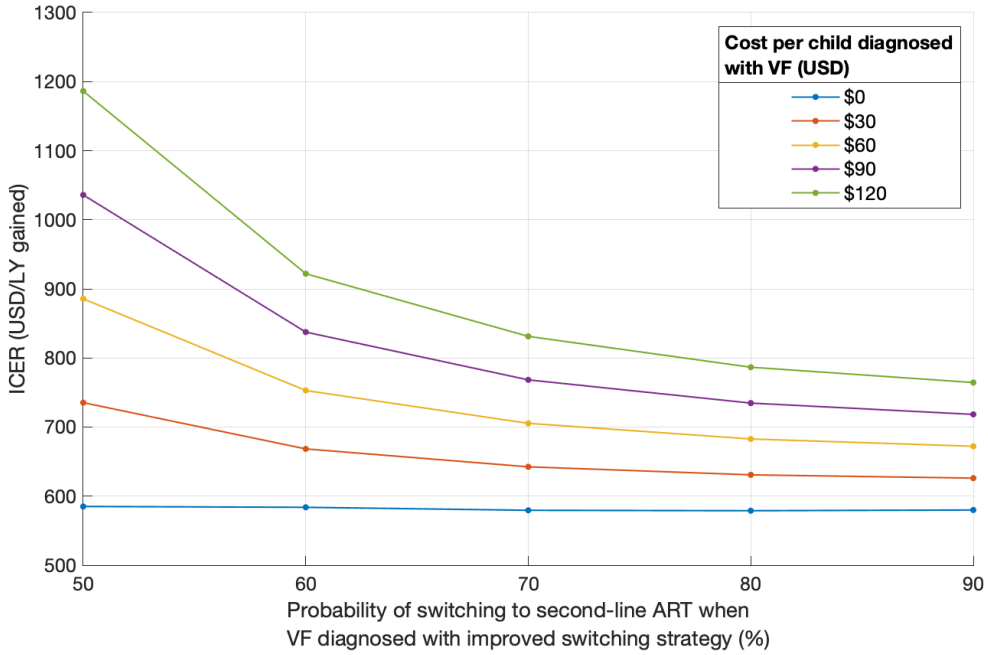


Figure 2-4. One-way sensitivity analyses of key model parameters.
A) The horizontal bars represent the range of ICER values of *improved switching* strategy compared to *status quo* from the tested parameters. The vertical bar represents the US$579/LY base-case analysis ICER. B) The horizontal bars represent the range of ICER values *DTG improved switching* strategy compared to *DTG status quo* from the tested parameters. The vertical bar represents the US$591/LY base-case analysis ICER.

A.



B.

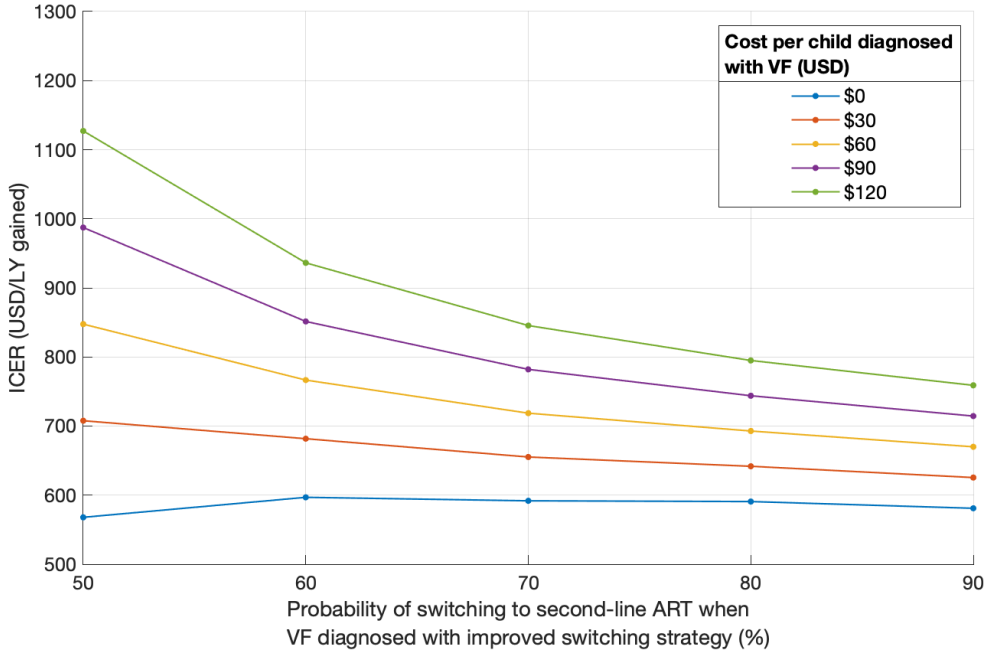


Figure 2-5. Scenario analyses with a cost incurred from improving the rate of switching to second-line ART.
(A.) ICER of *improved switching* strategy compared to *status quo* with the 40% base rate of switch resulted from varying the improved rate of switch from 50% to 80% and the cost per eligible person from $0 to $120 (B.) ICER of *DTG improved switching* strategy compared to *DTG status quo* with the 40% base rate of switch resulted from varying the improved rate of switch from 50% to 80% and the cost per eligible person from $0 to $120.

2.4. Discussion

With the availability of DTG, initiating NNRTI-based first-line ART in HIV-infected sub-Saharan African children aged ≥ 3 years in the presence of NNRTI-associated PDR may no longer be the most effective and cost-effective option. We evaluated alternative strategies in two settings: 1) In settings without DTG availability, we included the standard of care using NNRTI-based first-line ART, using the NNRTI-based first-line ART with improved switching to second-line PI-based ART, and PDR testing; 2) In the settings with DTG availability, we included a strategy using DTG-based first-line ART and DTG-based first-line ART with improved switching to second-line PI-based ART.

When DTG is not available, increasing the use of PI-based ART among children who were diagnosed with a virologic failure dominates PDR testing and is cost-effective based on the 2020 sub-Saharan Africa GDP of $850 (an ICER of $579/LY gained for *improved switching* compared to *status quo*). Additionally, combining DTG-based first-line ART with improving rate of switching has an ICER of $591/LY gained compared to initiating DTG-based first-line ART and is cost-effective. Comparing between strategies with NNRTI-based ART and DTG-based ART, we found that strategies with DTG provide higher benefit at lower costs. With greater access to DTG, HIV-infected children would benefit greatly from the antiretroviral drug that is more effective and less costly, especially in the setting with high PDR prevalence. Our results support the role of DTG as an initial treatment to HIV-infected children and provides a potential solution of improving a switch to second-line ART in the absence of DTG.

Our work extends upon prior studies on cost-effectiveness analyses of pediatric ART policies in sub-Saharan Africa [35,36] with the inclusion of DTG-based ART regimen as a first-line treatment for pediatric ART initiators. We also consider PDR testing as a guide in selecting an initial ART if DTG is not available, the cost-effectiveness of initiating ART from the age of 3 years, and incorporating the translation of CD4% to CD4 cell count at the age of 5 years. Cost-effectiveness analyses of adult ART initiator policies have found that initiating ART with DTG for all new initiators compared to NNRTI is likely to be cost saving [37] and PDR testing strategy is unlikely to be cost-effective compared to NNRTI use without testing [17]. In adult analyses prior to the availability of DTG, increasing the use of PI-based ART is potentially cost-effective compared to its best alternative [18]. Our findings are consistent with these adult studies, regardless of the differences between pediatric and adult HIV models, such as no HIV transmission, higher odds ratio for virologic failure of NNRTI, different mortality risks, and the uncertainties in our estimation of DTG-related parameters for HIV-infected children.

Our DTG-related parameters had to be estimated under several assumptions due to the lack of available information on DTG in HIV-infected children. We assumed the cost of DTG to be the same as the cost of NNRTI ($123 per person per year) and estimated the virologic failure rate of DTG in children from the adult data (9% after 1 year on ART). In 2017, there was a negotiation on pricing agreements for low- and middle-income countries to purchase DTG at the reduced price of $75 per person per year [16]. Assuming that the cost of DTG in children would be available at similar reduced price, the two strategies with DTG-based first-line ART would both be cost-saving, compared to the current standard of care.

With implementing new policies in mind, the cost-effectiveness of our strategies so far has been based on sub-Saharan Africa's 2020 GDP. However, there is an on-going debate on the threshold used for defining the cost-effectiveness of a health policy. Several studies have attempted to provide new thresholds [38–40]. The proposed thresholds for low- and middle-income countries include 4% to 51% of GDP per capita [39] and $500/DALY[2] averted [40]. We believe that whether a health policy is affordable, feasible, and cost-effective for a specific country would highly depend on several criteria such as the priority level of such disease, the country budget, opportunity cost, etc.

Our study has several limitations. First, we did not consider a longer time horizon in our analyses and, therefore, we do not provide the long-term benefits of dolutegravir. With increasing ART coverage and earlier ART initiating age, we believe that one of the long-term benefits of DTG would be an increase in viral suppression. We assumed that the monthly increase in CD4 in children on DTG is the same as those on NNRTI due to the lack of studies on children. In adults, DTG has been proven to provide a higher increase in CD4 than NNRTI [12,41]; if it has similar efficacy in children, we may be underestimating the gain in CD4. As higher CD4 is associated with lower risks of OI and mortality, the underestimated gain in CD4 leads to a possible overestimation of the base-case ICER in our analysis. Third, we assumed children failing a second-line ART continue staying on ART unless they were lost to follow-up. This assumption might not stay true in practice and would lead to an underestimation of ICER values. Lastly, we assumed the only available second-line ART is PI-based ART, although WHO recently suggested DTG-based ART as a potential regimen for second-line ART for children who failed NNRTI-based ART [11]. As there are currently very limited studies on DTG in children, we leave this area for future research.

In conclusion, improving the switch to second-line ART is a potential solution in the absence of DTG and is likely to remain cost-effective when DTG is available. Furthermore, initiating ART with dolutegravir in sub-Saharan African children will likely be effective and cost saving compared to the current standard of care. Our study provides evidence supporting the WHO updated HIV pediatric guidelines [11] and, suggests that the sooner children in resource-limited settings have full access to dolutegravir, the greater benefits they would gain.

[2] DALY: Disability-adjusted life year

3. Reinforcement learning models for personalized treatment selection in chronic depression

3.1. Introduction

Major depressive disorder (MDD) is one of the leading causes for disability and severely impacts quality of life in US adults [42]. In 2010, the economic burden of US adults with depression was estimated at $210 billion [43]. According to the National Center for Health Statistics (NCHS), approximately 7% of US adults experienced at least one major depressive episode in 2018 and 13% of US adults were reported using antidepressant at any given 30 day period during 2015 to 2018 [44]. For some patients, the duration of depressive symptoms can be at least 2 years or more, which is considered as chronic depression. Patients with chronic depression require treatment and monitoring for long periods as they tend to have lower and slower treatment response rates compared to those with nonchronic depression [45].

Depression patients typically proceed through three phases of treatment: acute phase (6-12 weeks), continuation phase (4 - 9 months), and maintenance phase (≥ 1 year) [45,46]. The aim of the acute phase is to achieve remission in a short period of time where the first choice of treatment could be antidepressant or depression-focused psychotherapy, or a combination of both depending on the depression severity [45,47,48]. According to clinical practice guidelines [46,47] and treatment guidelines for MDD [45], if patients do not respond or show partial response to antidepressant medication, the guidelines recommend to increase the medication dosage as the first option. If there is insufficient improvement, the options of switching to different medications or depression-focused psychotherapy should be considered. Finally, if patients still show no improvement, switching to different medications or the augmentation of psychotherapy should be considered. These clinical decisions are similar to decision rules, which are a series of if-then statements, where a certain action is chosen if a condition is met. After remission, patients proceed to the continuation phase for preventing relapse where they continue receiving the same treatment as in the acute phase. As patients with chronic depression have a higher risk of recurrence, they require treatment and monitoring for long periods. Patients with chronic depression are suggested to continue on the maintenance phase by receiving the same treatment as in the acute or continuation phase. If relapse or recurrence occurs, clinicians apply the same decision rules used in the acute phase.

The aforementioned rule-based treatment decisions, a one-size-fits-all rule, is commonly applied to patients with chronic depression. However, patient response to depression treatment varies greatly from patient-to-patient [49]. Due to this patient heterogeneity, a treatment plan that is tailored specific to each patient is needed to effectively improve the patient's health outcome [50]. Dynamic treatment regime estimation, a sequential decision-making problem under uncertainty, has been a research topic in the area of precision medicine with the focus on creating decision rules by leveraging patient data such as their health status, treatment preferences, personal history, family diseases, demographics, genetic, biomarker, and phenotypic characteristics [51,52]. Such information could be obtained from electronic medical records, personal reported information, and digital health technologies [53]. Traditionally, sequential decision-making problems under uncertainty that arise in the context of chronic diseases are commonly modeled as Markov Decision Processes (MDPs). For example, the problems of finding the optimal time to initiate a treatment were formulated as MDPs for patients with HIV [54], diabetes [55,56], and dialysis [57].

Another example is the screening problem for colorectal cancer [58] and prostate cancer [59], which were modeled as a Partially Observable Markov Decision Process (POMDP) – a subcategory of MDP. The design of personalized treatment strategies for breast cancer [60] and hypertension [61] were formulated as MDPs. To solve MDPs with finite state and action spaces and known transition probabilities and reward functions, a classical approach is dynamic programming [62]. However, problems with a large state-space (large-scale MDPs) fall under the curse of dimensionality and dynamic programming becomes intractable. Consequently, a more efficient approach is required for solving such problems. In contrast to the model-based approach, the model-free approach is flexible as it does not require knowledge of the disease model. One of the approaches is reinforcement learning (RL), sometimes referred to as approximate dynamic programming and can provide a near-optimal policy to large-scale and complex MDPs [63].

RL has been applied to personalized medicine for chronic diseases, which assists in making decisions on the selection of possible treatment strategies and the amount of dosage for drugs. Several studies have applied RL techniques to estimate optimal treatment regime on various diseases such as depression, HIV, and cancer [64–67]. Q-learning, a model-free approach, is one of the common off-policy methods that use pre-collected data in learning. Q-learning is a temporal difference algorithm where the action-value function or Q-function is updated based on a single observed state-action pair at a time [63]. Several studies have applied Q-learning to design personalized treatment strategy for radiotherapy [68] and for patients with non-small cell lung cancer [67]. Similar to Q-learning, Fitted Q-iteration (FQI) also learns the action-value function, however, it is a batch-mode RL algorithm. Instead of using a single state-action pair, FQI uses all observed state-action pairs in the updating process, and therefore, converges faster than Q-learning [69]. FQI was applied to create adaptive treatment strategies for patients with epilepsy [70] and type 1 diabetes [71]. Escandell-Montero et al. [72] have shown that FQI algorithm outperforms Q-learning in the dosage management problem.

In the context of mental illness, there exists several studies on adaptive treatment strategies that apply both Q-learning using depression data from the STAR*D trial[3] [64,73–76] and FQI to data from CATIE[4] [77]. However, both trials have short time horizons with few decision points, e.g., STAR*D has no more than 7 decision points and CATIE has only 2 decision points. Shortreed et al. [77] used a linear function for fitting Q-functions for their FQI. Other function approximators for FQI in other studies include but are not limited to kernel regression [71], random forest regression [71], and tree-based regression (extremely randomized trees) [59,70,78]. To our knowledge, FQI has not been used to design chronic depression treatment strategies with many decision points over a long-time horizon.

To assist mental health professionals in decision-making, this study aims to design personalized treatment algorithms for patients with chronic depression in the maintenance phase. We formulated a stylized treatment selection problem (a sequential decision-making problem), as an MDP with the objective of finding the optimal treatment policy for different type of patients, which are categorized by their characteristics of disease progression. We assumed there are three available treatments that have different levels of effectiveness, which is represented by a treatment effect

[3] STAR*D: Sequenced Treatment Alternatives to Relieve Depression trial

[4] CATIE: Clinical Antipsychotic Trials of Intervention Effectiveness is a multistage clinical trial of patients with schizophrenia.

that is unique to each patient and each treatment. We assumed each patient responds best to one of the treatments, and the ideal policy is to give the most effective treatment to the right patient. To find the optimal policy, we developed three heuristic approaches: 1) a rule-based policy modeled after the STAR*D (Human Intuition) [79], 2) a model-based approach as a variation of one-step look-ahead policy (Update), and 3) a model-free approach such as a value-based reinforcement learning (FQI).

Previous studies in personalized treatment have studied the performance among model-free approaches, i.e., variants of Q-learning vs. traditional Q-learning [80], FQI vs. Q-learning [72], or variants of FQI vs. FQI [70]. However, it is unclear whether the model-free approaches perform better than the model-based approaches. Therefore, it is important to compare the model-based and model-free approaches' policy performances as alternatives to the rule-based policy, and evaluate their ability to identify the most effective treatments for each patient with many decision time points.

This chapter is organized as follows. Section 3.2 introduces the methodologies used in formulating our problem as a MDP, the definition of treatment effect, and the three heuristic approaches. Section 3.3 describes a computer experiment using simulated depression trajectories to compare the three proposed approaches. We evaluate their policy performances using simulated patient health outcomes and accuracy in treatment selection. These methods are also evaluated under experimental setups that have various levels of patient heterogeneity in treatment effect and disease progression. In Section 3.4, we present the results and evaluate robustness of the policies by conducting sensitivity analyses on parameters affecting health outcomes. Lastly, we conclude with a discussion of our results, methods, and limitation in Section 3.5.

3.2. Methodology

3.2.1. Problem formulation

We considered a treatment selection problem for patients with chronic depression, which is a sequential decision-making problem with finite states and finite time horizon. The objective is to maximize the patient health outcomes by finding the optimal sequences of treatment using an optimal treatment policy. We assumed there is a holding period after every treatment switch. A holding period refers to a duration of several time periods for patients being on the same treatment where their responses can be observed. No switch in treatment is allowed unless a patient is intolerant to a treatment. We formulated the problem under a MDP framework using the following definitions:

Time Period: Let t denote a time period in which a state is observed, where $t = 1, 2, \cdots, T$. For chronic depression, the Patient Health Questionnaire (PHQ)-9 [81] is a questionnaire used to measure depression severity by asking patients about their symptoms in the past two weeks. Therefore, we considered a bi-weekly period in our problem.

Time Horizon: Let T denote the time horizon of the treatment selection problem. Note that we considered a possibility of death during patients' disease progression. If death occurs before period T, a state of death is observed in every period thereafter and no decision is required.

States: Our state s consists of health state and treatment state; $s = (h, m)$. Let h_t denote a health state observed at time t and let H denote a state space of h_t. Let m_t denote a treatment state, which represents a treatment patient received in period t and M denote the state space of m_t. We considered a finite set of discrete health states and a finite set of treatments. The health state is defined by the depression severity level and death; $H = \{1, 2, 3, \text{death}\}$. The definition of each level is as follows: level 1 refers to 'mild', level 2 refers to 'moderate', and level 3 refers to 'severe' levels of depression. We considered three treatment states; $\text{M} = \{\text{treatment 1, treatment 2, treatment 3}\}$, where treatment 3 is the most effective treatment and treatment 1 is the least effective treatment observed by population average. Each treatment is assumed to have side effects where more effective treatment is associated with increased side effects.

Actions: At any period t, an action is a selection of available treatments if $h_t \neq death$. Let a_t denote the action taken at period t and is implemented in period $(t + 1)$, therefore, $a_t = m_{t+1}$. The action state space A includes {treatment 1, treatment 2, treatment 3}.

Transition probability: A transition from the current state s_t to the next state s_{t+1} given the chosen action a_t is associated with a probability of $p(s_{t+1}|s_t, a_t)$. We modeled a patient's chronic depression progression using a Markov model (Figure 3-1). For convenience, we denote $p(h_{t+1}|h_t)$ as a non-treatment transition probability from health state h_t to state h_{t+1} where $\sum_{h_{t+1}} p(h_{t+1}|h_t) = 1$ for all $h_t \in H$. We denote $p^{m_{t+1}}(h_{t+1}|h_t)$ as a treatment transition probability from health state h_t to state h_{t+1}while being on treatment m_{t+1} in period $(t + 1)$ where $\sum_{h_{t+1}} p^{m_{t+1}}(h_{t+1}|h_t) = 1$ for all $h_t \in H, m_{t+1} \in M$.

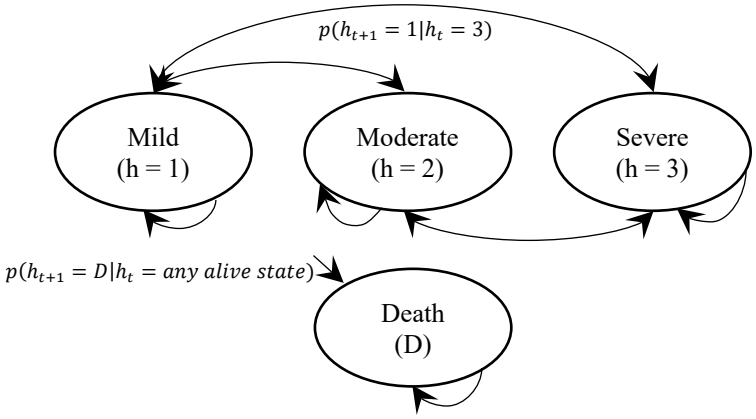


Figure 3-1. Chronic depression health state-transition diagram

All alive health states can transition among themselves with transition probabilities, e.g. a patient with current health state of 'level 1' can transition to health state 'level 3' with a non-treatment transition probability of $p(h_{t+1} = 1|h_t = 3)$ if not on treatment. If a patient is on treatment m_t, the treatment transition probability is $p^{m_t}(h_{t+1} = 1|h_t = 3)$. All alive states can transition to death, which is an absorbing state.

Immediate Reward: Immediate reward $r_{t+1}(s_{t+1}|s_t, a_t)$ is a reward obtained from a transition of the current state s_t to the next state s_{t+1} as a result of the chosen action a_t. In our problem, the immediate reward at time $t+1$ is the utility of living in a health state h_{t+1}; $u(h_{t+1})$ while being on a treatment m_{t+1}, which is associated with side effects (disutility, $d(m_{t+1})$); $r_{t+1}(s_{t+1}|s_t, a_t) = u(h_{t+1}) \times \{1 - d(m_{t+1})\}$.

Policy: A policy is a decision rule that maps from state to action. Let $\pi(s_t)$ denote a policy for state s_t where $\pi(s_t) = a_t$ for $a_t \in A$ and let π^* denote the optimal policy, which is a policy that provides rewards that are higher than or equal to other policies.

Value function: A value function $V_\pi(s)$ is the expected reward under policy π when starting in state s. For all $s = (h, m)$ where $h \in H$ and $m \in M$,

$$V_\pi(s) = E_\pi\left[\sum_{t=0}^{T-1} \gamma^t r_{t+1}(s_{t+1}|s_t, \pi(s_t)) + \gamma^T g(s_{T+1}) \middle| s_t = s\right]$$

where γ is a discount factor and $g(s_{T+1})$ is a terminal reward from being in state s_{T+1}.

The optimal value function $V^*(s)$ denotes the value function under the optimal policy and can be expressed as:

$$V^*(s) = \max_\pi V_\pi(s)$$

Q-function: Q-function $Q_\pi(s, a)$ or action-value function is the expected reward under policy π when starting in state s and the chosen action is a. For all $s = (h, m)$ where $h \in H$ and $m \in M$ and for all $a \in A$,

$$Q_\pi(s, a) = E_\pi[r_{t+1}(s_{t+1}|s_t, a_t) + \gamma V_\pi(s_{t+1})|s_t = s, a_t = a]$$

The optimal Q-function $Q^*(s, a)$ denotes the action-value function under the optimal policy and can be expressed as:

$$Q^*(s, a) = \max_\pi Q_\pi(s, a)$$

Therefore, the relationship between $V^*(s)$ and $Q^*(s, a)$ can be expressed as:

$$V^*(s) = \max_{a \in A} Q_{\pi^*}(s, a)$$

3.2.2. Treatment effects

In this section, we explain how treatment effects are modeled in this study. Treatment can improve a patient's disease progression, and we simulated this improvement by incorporating a treatment effect to the patient's non-treatment transition matrix. We assumed that each patient may have their own response to each available treatment, which is represented by a treatment effect ρ^m. Once patients are on treatment, their progression is described by a treatment transition matrix P^m, while a non-treatment transition matrix P is used for untreated patients. To obtain P^m, ρ^m is incorporated into a non-treatment transition matrix P based on the following assumptions.

Let ρ^m denote a treatment effect by any treatment m where $\rho^m \in [0,1]$.
Let P denote a non-treatment transition matrix and $p(h_{t+1}|h_t)$ is an element of P.

$$P = \begin{bmatrix} p(1|1) & p(2|1) & p(3|1) & p(D|1) \\ p(1|2) & p(2|2) & p(3|2) & p(D|2) \\ p(1|3) & p(2|3) & p(3|3) & p(D|3) \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Let P^m denote a treatment transition matrix and $p^m(h_{t+1}|h_t)$ is an element of P^m.

$$P^m = \begin{bmatrix} p^m(1|1) & p^m(2|1) & p^m(3|1) & p(D|1) \\ p^m(1|2) & p^m(2|2) & p^m(3|3) & p(D|2) \\ p^m(1|3) & p^m(2|3) & p^m(3|3) & p(D|3) \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

We assumed that any treatment would improve the chance of transitioning to better health states (solid line box) and reduce the chance of transitioning to worse health states (unboxed elements). Therefore, we multiply the probabilities of transitioning to worse health states by the treatment effect ρ^m. The probabilities of transitioning to death in P^m (dashed line box) are assumed to be the same as P. Thus, the probabilities of transitioning to better health states (solid line box) are calculated by subtracting 1 with the probability of transition to death and the reduced probability of transitioning to worse health states.

Therefore, P^m can be expressed as $P^m = PA + B$ where A and B are as follows:

$$A = \begin{bmatrix} \rho^m & 0 & 0 & 0 \\ 0 & \rho^m & 0 & 0 \\ 0 & 0 & \rho^m & 0 \\ 0 & 0 & 0 & \rho^m \end{bmatrix}$$

$$B = \begin{bmatrix} q_{11}\{1-\rho^m\}\{1-p(D|1)\} & 0 & 0 & p(D|1) - \rho^m p(D|1) \\ q_{21}\{1-\rho^m\}\{1-p(D|2)\} & q_{22}\{1-\rho^m\}\{1-p(D|2)\} & 0 & p(D|2) - \rho^m p(D|2) \\ q_{31}\{1-\rho^m\}\{1-p(D|3)\} & q_{32}\{1-\rho^m\}\{1-p(D|3)\} & 0 & p(D|3) - \rho^m p(D|3) \\ 0 & 0 & 0 & p(D|D) - \rho^m p(D|D) \end{bmatrix}$$

where

$q_{uv} = \frac{p(v|u)}{\sum_{v \le u, v \neq Z} p(v|u)}$ and Z is the worst alive health state, which is health state 3 in this case.

Proposition 1 for P^m:

1. The probabilities of transitioning to better health states (solid line box) are higher than those of P.
2. The probabilities of transitioning to worse health states excluding the probabilities transition to death (unboxed elements) are lower than those of P.

The proof for Proposition 1 is provided in the Appendix B-I.1.

3.2.3. Heuristic algorithms

Next, we provide a detailed explanation of the three heuristic algorithms proposed for creating personalized treatment plans, namely, Human intuition algorithm, One-step look-ahead with estimated treatment effects (Update), and Fitted Q-Iteration (FQI). The objective is to maximize health outcomes. Human intuition algorithm is composed of decision rules that are similar to how a mental healthcare provider would select treatment. Update algorithm is a model-based algorithm, where the full knowledge of disease progression prior to treatment is known. And in contrast, FQI is a model-free algorithm in which the policy is obtained through learning from patient treatment experiences.

Comparing model-based and model-free algorithms, one would expect that the model-based algorithms would perform better as the disease progression prior to treatment is known. However, in the presence of heterogeneity, the exact information on individual disease progression or individual treatment effect is unlikely to be acquired. It is debatable whether model-based or model-free algorithms perform better in this regard. Therefore, we investigated the performance of the two algorithms using simulation setups under two levels of heterogeneity.

To measure the algorithm performance, we generated a set of patient trajectories by the treatment policy created from each algorithm and use health outcome as the performance metric. The health outcome is calculated over a patient's trajectory, which is a summation of health utility scores of the sequence of health states and treatments. The health outcome can be expressed as:

Let $\mathcal{T}$ denote a health trajectory of patient j with a length of T and $\mathcal{T} = (s_1, \dots s_T)$.
Let $R_j(\mathcal{T})$ denote the health outcome of a trajectory of patient j.

$$R_j(\mathcal{T}) = \sum_{t=1}^{T} u(h_t) \times \{1 - d(m_t)\}$$

1) Human intuition algorithm

To imitate real-life practices of how medical practitioners would select treatment, we created a rule-based policy in selecting treatment based on the rules used in the STAR*D trial [79]. The rules of selecting a treatment are as follows:

Let O_t denote an observational sequence of states from a fixed follow-up duration of L period (i.e. the holding period) at time period t and $O_t = \{s_{t-L-1}, s_{t-L-2}, \dots, s_{t-1}, s_t\}$ where $s_t = (h_t, m_t)$. The rules are:

a. If h_t is 'level 1' and 'level 1' is in $O_t \backslash \{h_t\}$, then continue on the same treatment.
b. If h_t is worse than h_{t-L-1} and 'level 1' is not in O_t, then select the most effective treatment if $m_t \neq$ the most effective treatment.
c. If 'level 3' appears in O_t more than 3 times, then select the most effective treatment if $m_t =$ the most effective treatment.
d. If h_t is better than h_{t-L-1}, then randomly select among treatment 1 and treatment 2.
e. Otherwise, randomly select a different treatment than the current treatment.

2) One-step look-ahead with estimated treatment effects—Update

Update is a model-based greedy policy with two main assumptions. First, we have a full knowledge of the non-treatment transition matrix. Second, the method of incorporating the treatment effect to a non-treatment transition matrix is known. In this greedy policy, the action is selected by maximizing the expected reward received in the next period. The optimal action a_t^* is formulated as follows:

Let a_t^* denote the optimal action at period t.
Let r_{t+1} denote a reward received at period $t+1$ from being in state s_{t+1} where $s_{t+1} = (h_{t+1}, m_{t+1})$.
Let $\hat{p}^{m_{t+1}}(h_{t+1}|h_t = x)$ denote an estimated treatment transition probability from a known health state h_t of x to health state h_{t+1}.

$$a_t^* = \underset{m_{t+1}}{\operatorname{argmax}} E[r_{t+1}(s_{t+1})] = \underset{m_{t+1}}{\operatorname{argmax}} E[r_{t+1}(h_{t+1}, m_{t+1})] \quad (1)$$

$$E[r_{t+1}(h_{t+1}, m_{t+1})] = \sum_{h_{t+1}} \hat{p}^{m_{t+1}}(h_{t+1}|h_t = x) \times u(h_{t+1}) \times \{1 - d(m_{t+1})\} \quad (2)$$

where $\hat{p}^{m_{t+1}}(h_{t+1}|h_t = x)$ is obtained from incorporating a moving average of estimated treatment effect $\hat{\rho}^{m_{t+1}}$ to a non-treatment transition matrix P. $\hat{\rho}^{m_{t+1}}$ is estimated only after a holding period.

For any treatment m, a treatment effect $\hat{\rho}^m$ is estimated using the maximum likelihood estimation or MLE. For convenience, we present the log-likelihood function using the notation ρ for treatment effect. Let O denote an observational sequence from a holding period with a length of L for any time t, $O = (s_{t-L-1}, s_{t-L-2}, \dots, s_{t-1}, s_t)$. The probability of observing a sequence O is $p(O) = \prod_{x_1,x_2} p^m(x_1, x_2)^{N(x_1,x_2)}$ where $N(x_1, x_2)$ denotes the number of transitions from x_1 to x_2; $x_1, x_2 \in H$ and $p^m(x_1, x_2)$ denotes an element of transition $p^m(h_{t+1} = x_2|h_t = x_1)$ of a treatment transition matrix P^m where treatment m is a treatment taken during a holding period; $m = m_t, \forall t \in \{t-L-1, t-L-2, \dots, t-1, t\}$. The log-likelihood function of $p(O)$ is expressed as follows:

$$\begin{aligned}\log p(O) &= \sum_{x_1}\sum_{x_2} N(x_1, x_2) \log p^m(x_1, x_2) \\ &= \sum_{x_1 \neq D} \Bigg\{ N(x_1, D)\log p(x_1, D) + \sum_{x_2 > x_1} N(x_1, x_2)\log \rho p(x_1, x_2) \\ &\qquad + \sum_{x_2 \leq x_1, x_2 \neq 3} N(x_1, x_2)\log\left[\rho p(x_1, x_2) + q_{x_1 x_2}\{1-\rho\}\{1 - p(x_1, D)\}\right] \Bigg\} + N(3,3)\log \rho p(3,3)\end{aligned}$$

where $p(x_1, x_2)$ denotes an element of transition $p(h_{t+1} = x_2|h_t = x_1)$ of a non-treatment transition matrix P.

Let $f(\rho)$ denote $\log p(O)$ and the maximum likelihood estimator $\hat{\rho}$ is a solution of the following maximization problem:

$$\max_{\rho} \; f(\rho)$$
$$\text{subject to} \quad \bar{\rho} - k\sigma \leq \rho \leq \bar{\rho} + k\sigma \tag{3}$$

where $\bar{\rho}$ denotes the population average treatment effect, σ denotes the standard deviation of the treatment effect, and k denotes a constant number. If $\bar{\rho}$ and σ are unknown, we use 0 and 1 as the lower and upper bound of ρ.

We presented the solution to the above maximization problem using Lagrange multipliers in Appendix B-I.2. The optimal solutions ρ^* cannot always be expressed as a close-form solution. Therefore, we prove that $f(\rho)$ is a concave function (Appendix B-I.3) and we used CVXPY[5] to solve for ρ.

3) Fitted Q-Iteration (FQI)

FQI [69] is a batch-mode model-free reinforcement learning technique that aims to estimate the Q-function. In FQI, the algorithm updates the Q-function using a complete dataset of state-action pairs at each updating iteration. We considered two models in approximating the Q-function: 1) Ridge regression (FQI-Ridge) and 2) Random forest regression (FQI-RF). Ridge regression is a linear approximator, which is a variation of a linear regression. The difference between ridge regression and linear regression is an inclusion of a penalty term or the L2-norm of the coefficients. Random forest regression is a nonlinear approximator, it is an ensemble model that creates many parallel decision trees and provides a mean of all trees as a final prediction. Next, we explained the FQI algorithm under our problem setting using patient trajectories as input data.

We first considered an existing dataset of N trajectories where each trajectory is from a single patient and has a length of l periods. Each trajectory contains a sequence of our MDP states; health states and treatment states $(s_t)_{t=1}^{l}$ where $s_t = (h_t, m_t)$. To learn the Q-function via regression models, we considered patients' responses during a holding period in addition to the most recent state s_t. This is consistent with our model-based algorithm which utilizes patients' responses from a holding period to estimate treatment effect. Therefore, we defined a state of the Q-function as a feature vector x_t that consists of the patient's observational history on health and treatment states collected from period $t-5$ to period t. The feature vector x_t is defined in Table 3-1. For convenience, we created a dataset $\mathcal{D}$ of transition tuples $\{(x_t^i, a_t^i, x_{t+1}^i, r_{t+1}^i)_{i=1}^{|\mathcal{D}|}\}$ aggregated from all N trajectories where index i denotes a transition i in any given trajectory and the number of transition tuples is $|\mathcal{D}|$[6].

[5] CVXPY is a modeling language for construction and solving convex optimization problems for Python.

[6] For our problem, the number of transition tuples depends on the dimension of x_t and the length of each trajectory l. Therefore, we define the number of transition tuples in $\mathcal{D}$ as $|\mathcal{D}|$. For example, if a single trajectory results in 5 transition tuples. Then, a dataset $\mathcal{D}$ has a total of $|\mathcal{D}| = 5N$ transition tuples.

Let L denote a duration of a holding period.

Let $\hat{Q}_k(x, a)$ denote an approximated Q-function at iteration k for any inputs of x and a.

1. Create a dataset set $\mathcal{D} = \{(x_t^i, a_t^i, x_{t+1}^i, r_{t+1}^i)_{i=1}^{|\mathcal{D}|}\}$ from N trajectories where $a_t = m_{t+1}$ and $r_{t+1} = \sum_{j=t+1-L}^{t+1} u(h_j) \times \{1 - d(m_j)\}$. x_t and x_{t+1} are defined in Table 3-1.
2. Initialize $\hat{Q}_0(x_t^i, a_t^i) = r_{t+1}^i$ for all $i = 1, \ldots, |\mathcal{D}|$
3. Let $k = 1$
4. Repeat the steps below until k reaches the maximum number of K iterations:
    a. Based on a dataset $\mathcal{D}$ and $\hat{Q}_k$, create a training dataset of the inputs $(x_t^i, a_t^i)_i$ and targets $\hat{Q}_{k-1}(x_t^i, a_t^i)$ for $i = 1, \ldots, |D|$
    b. Approximate $\hat{Q}_{k-1}$ using supervised learning (in our case either ridge regression or random forest regression) with inputs and targets from a.
    c. Update $\hat{Q}_k$ with Bellman equation:
    $$\hat{Q}_k(x_t^i, a_t^i) = r_{t+1}^i + \gamma \max_{a'} \hat{Q}_{k-1}(x_{t+1}^i, a')$$
    where γ denotes a discounting factor.
    d. Calculate the distance between $\hat{Q}_k(x, a)$ and $\hat{Q}_{k-1}(x, a)$ to check for convergence
    e. $k = k + 1$

Table 3-1. The state x_t components for ridge and random forest regressors

State component definition	x_t	Data type
Health states	h_{t-5}	Categorical
	h_{t-4}	
	h_{t-3}	
	h_{t-2}	
	h_{t-1}	
	h_t	
Treatment states	m_{t-5}	Categorical
	m_{t-4}	
	m_{t-3}	
	m_{t-2}	
	m_{t-1}	
	m_t	
Interaction between health state and treatment state	h_{t-5}*m_{t-5}	Categorical of all possible combinations of h_t and m_t
	h_{t-4}*m_{t-4}	
	h_{t-3}*m_{t-3}	
	h_{t-2}*m_{t-2}	
	h_{t-1}*m_{t-1}	
	h_t*m_t	
The number of times each alive health state appears in a trajectory	$n_{h=1}$	Numerical
	$n_{h=2}$	
	$n_{h=3}$	

3.3. Simulation experiments

To evaluate the three algorithms, we used simulated patient trajectories. We simulated three groups of patients that can be categorized by their progression pattern: fast degrading, slow degrading, and steady progression. For each progression group, we divided 120 patients evenly into three treatment types; each type has the best treatment response on one of the three treatments. The holding period is assumed to be 6 bi-weekly periods, which is equivalent to the follow-up period in the STAR*D trial, unless a dropout due to intolerance occurs. The time horizon of the problem is 120 bi-weekly periods.

We assumed that all patients start with a health state of 'level 3' or severe, similar to the patients in the STAR*D trial. According to a cost-effectiveness study on antidepressants and cognitive behavioral therapy [82], patients diagnosed with major depressive disorder were initiated with either antidepressant only or cognitive behavioral therapy only. Therefore, we randomly select the initial treatment among treatment 1 (cognitive behavioral therapy or CBT only) and treatment 2 (antidepressant only). Treatment 3 is a combination of treatment 1 and 2. Note that we considered each treatment as a treatment type – not a specific treatment.

Each treatment is assumed to be associated with side-effects, which are reflected by the disutility score. Additionally, at every holding period after the patient switches to a different type of treatment, there is a chance that the patient may be intolerable to the treatment at the 4^{th} period of the holding period (week 8 of being on treatment [47]). For example, suppose a patient switches to treatment 2 and is in a holding period. If this patient is intolerable to treatment 2, he/she drops out of the current line of antidepressant. Then, a new treatment will be selected from the three available treatment types where treatment 2 is now a new line of antidepressant.

To evaluate robustness of the algorithms' performance toward the heterogeneity, we considered two experimental setups that correspond to low heterogeneity and high heterogeneity. The two setups and their assumptions are as follows:

1) Setup 1: Heterogeneity in treatment response type
   Treatment response types are unknown.
   Patients respond to the same treatment have same values of treatment effects.

2) Setup 2: Heterogeneity in treatment response type and treatment effects
   Treatment response types are unknown.
   Patients respond to the same treatment have different values of treatment effects but the ranks of treatment effects for the three treatments are the same.

In the following subsections, we present a detailed description of patient progression groups, treatment response types, and treatment effects for the experiments. The simulation of patient trajectories and how Update and FQI algorithms were applied to the simulated trajectories are explained in full detail.

3.3.1. Patient disease progression as a transition matrix

Each progression group i is represented by a canonical non-treatment transition matrix, P_i. The three canonical matrices representing bi-weekly disease transition are presented Table 3-2.

Table 3-2. Canonical transition matrices by progression pattern group

Progression	Slow degrading (P_1)	Fast degrading (P_2)	Steady (P_3)
Canonical Matrix	$\begin{bmatrix} 0.6 & 0.3 & 0.1 & \varepsilon \\ 0.1 & 0.6 & 0.3 & \varepsilon \\ 0.1 & 0.3 & 0.6 & \varepsilon \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.2 & 0.7 & 0.1 & \varepsilon \\ 0.1 & 0.2 & 0.7 & \varepsilon \\ 0.08 & 0.16 & 0.76 & \varepsilon \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.05 & 0.05 & \varepsilon \\ 0.05 & 0.9 & 0.05 & \varepsilon \\ 0.05 & 0.05 & 0.9 & \varepsilon \\ 0 & 0 & 0 & 1 \end{bmatrix}$

where $\varepsilon \approx 0.00003$, a small bi-weekly probability of transitioning to death. We calculated this probability by multiplying 1.58, the relative risk of mortality in patients with depression compared to patients without depression [83], with a bi-weekly death probability converted from the 2017 US life tables [84].

3.3.2. Treatment effects simulation

We categorized patients into three treatment response types: 1) responds to treatment 1, 2) responds to treatment 2, and 3) responds to treatment 3. We assumed treatment 3 is the most effective and treatment 1 is the least effect, by population average.

In Setup 1, we assumed that patients who respond to the same treatment have the same treatment effect values for all three treatments (Table 3-3). Since we assumed evenly distributed treatment response types, the averages of the treatment effects are 0.72, 0.65, and 0.58 for treatment 1, treatment 2, and treatment 3, respectively. A smaller treatment effect value represents a more effective treatment.

Table 3-3. Treatment effects by response types for setup 1

Response type	Treatment 1	Treatment 2	Treatment 3
Respond to treatment 1	0.45	0.85	0.65
Respond to treatment 2	0.85	0.45	0.65
Respond to treatment 3	0.85	0.65	0.45
Population average	0.72	0.65	0.58

In Setup 2, we assumed that the underlying distribution for treatment effects of each treatment is a Beta distribution (Figure 3-2). The mean effects of treatment 1, treatment 2, and treatment 3 are 0.75, 0.65, and 0.55, respectively, with standard deviations of 0.05, 0.08, and 0.07.

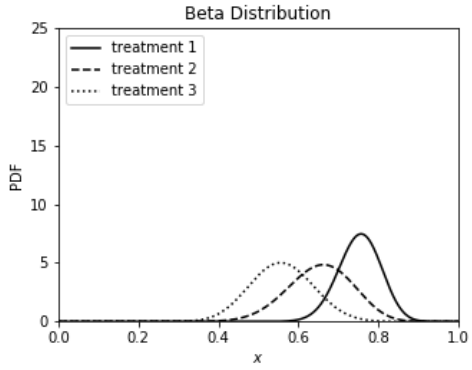


Figure 3-2. The Beta distributions of treatment effects for three treatments

We sampled the treatment effects from the underlying Beta distributions with the lower and upper bounds specified according to the patient response type in Table 3-4. Note that we specified the following lower and upper bounds such that the sample average of the three treatments across all response types are similar in values to the sample average in Setup 1 and the population average. The resulting sample average of treatment 1, treatment 2, and treatment 3 in Setup 2 are 0.72, 0.62, and 0.57, respectively.

Table 3-4. The lower and upper bounds for treatment effects sampling setup 2

Response type	Treatment 1	Treatment 2	Treatment 3
Respond to treatment 1	[0.54, 0.59]	[0.64, 0.89]	[0.59, 0.79]
Respond to treatment 2	[0.75, 0.91]	[0.41, 0.46]	[0.59, 0.79]
Respond to treatment 3	[0.75, 0.91]	[0.64, 0.89]	[0.40, 0.45]

3.3.3. Trajectory simulation

To simulate a patient trajectory using each algorithm, we apply the following steps:

For $t = 1, \dots, T$

1. Initialize a patient health state with 'level 3' ($h_t = 3$) and randomly select an initial treatment m_t among treatment 1 and treatment 2.
2. Generate a health sequence for a holding period (6 periods)
   For $j = 1, \dots, 6$
   a. Generate h_{t+1} by randomly sampling from H with a probability distribution $p^{m_{t+1}}(h_{t+1}|h_t)$ obtained from $P^{m_{t+1}}$.
   b. If $j = 4$, check if a patient is intolerable to the current treatment. If the treatment is intolerable, go to step 3 immediately. Otherwise, continue until a holding period is completed (j = 6).

3. Update t and choose action based on the algorithm

$$a_t = \begin{cases} \text{rules by 'human intuition'}, & \text{if algorithm is Human intuition} \\ \underset{m_{t+1}}{\text{argmax}}\, E[r_{t+1}(h_{t+1}, m_{t+1})], & \text{if algorithm is Update} \\ \underset{a'}{\text{argmax}}\, \hat{Q}(x_t, a'), & \text{if algorithm is FQI} \end{cases}$$

4. Let $m_{t+1} = a_t$
    a. If $m_{t+1} \neq m_t$, then go back to step 2 (a treatment switch generates a new holding period).
    b. If $m_{t+1} = m_t$, then generate the next health state h_{t+1} and go back to step 3.

3.3.4. Learning process for FQI and Update algorithms

For both setups (setup 1 and setup 2), FQI algorithm learns a policy from a set of 120 patient trajectories obtained from a trial with a duration of 12 periods. The learned policy is then applied to the same set of patients to generate their continued trajectories in the next trial. This process was repeated until the maximum number of trials was reached (10 trials). We assumed all 120 patients are from the same known progression group. We considered 120 patients that are equally distributed into three response types (40 each). The learning process is as follows:

1) Generate a set of 120 trajectories for 120 patients using our human intuition algorithm where each trajectory has 12 periods or 6 months with 'level 3' as the first health state and 'treatment 1' or 'treatment 2' as the first treatment. Let $\mathcal{D}_1$ denote the first dataset of 120 trajectories where each has 12 periods.
2) For trial $c = 1, \dots, 10$
    a. Use a dataset $\mathcal{D}_c$ to do the following tasks:
        a. For Update algorithm, initialize treatment effect $\widehat{\rho_j}^m$ using $\mathcal{D}_c$ with MLE for all $j = 1, \dots, 120$ and $m \in M$.
        b. For FQI algorithm, train the Q-function $\hat{Q}_K^c$ with $\mathcal{D}_c$ for a maximum training iteration $K = 50$ and a discount factor γ of 0.8.
    b. Generate the next set of 120 trajectories $\mathcal{T}_c$ with a length of 12 periods for the same 120 patients, which is a continuation of trajectories from $\mathcal{D}_c$. The treatment selection within a generated trajectory for each algorithm is as follows:
        a. For Update algorithm, select the treatment based on Eq. (1) and (2) using the estimated $\widehat{\rho_j}^m$, which is updated with MLE after every 6 bi-weekly holding period.
        b. For FQI algorithm, select the treatment based on $\hat{Q}_K^c$.
    c. For both algorithms, create a new dataset $\mathcal{D}_{c+1}$ where $\mathcal{D}_{c+1} = D_c \cup \mathcal{T}_c$. Therefore, each trajectory in $\mathcal{D}_{c+1}$ is twice the length of those in $\mathcal{D}_c$.
    d. Let $c = c + 1$, and repeat step 2a.

For FQI algorithms, the Q-function is approximated via experience learning. Therefore, the Q-function needs to converge during the learning process. After convergence, we apply the learned policy obtained from the approximated Q-function to the simulated patients. The convergence plots of FQI-Ridge and FQI-RF are presented in Appendix B-II.2.

3.3.5. Parameters

We calculated the health outcomes of the 120 trajectories from each algorithm using the parameter values presented in Table 3-5. We also conducted sensitivity analyses using a combination of parameter ranges.

Table 3-5. Parameter values for simulation experiments

	Base case	Range	Reference
Utility by health state			
Level 1 (Mild)	0.70	0.67– 0.73	[85]
Level 2 (Moderate)	0.52	0.49– 0.56	
Level 3 (Severe)	0.39	0.35 – 0.43	
Death	0		
Disutility by treatment			
Treatment 1 (CBT)	0.01		[86]
Treatment 2 (Antidepressant)	0.06	0.04 – 0.10	[87]
Treatment 3 (CBT + Antidepressant)	0.07		Assumed as a summation of treatment 1 and treatment 2
Probability of dropout at week 8 during holding period			
Treatment 1 (CBT)	5.52%	RR of dropout, CBT to antidepressant: 0.06 – 2.50 (base-case: 0.40)	RR from Ross et al. 2019 [82]
Treatment 2 (Antidepressant)	13.8%	11.2 – 16.3%	[88]
Treatment 3 (CBT + Antidepressant)	13.8%	11.2 – 16.3%	Assumed the same as treatment 2

3.4. Results

Four algorithms were applied to both experimental setups: Human intuition (a rule-based approach), Update (a model-based approach), FQI-Ridge and FQI-RF (model-free approaches). Both FQI algorithms converged within 50 training iterations in our experiments. The final models of Q-function approximator by FQI-Ridge and FQI-RF are presented in Appendix B-II.1. We evaluated the performance of the policies derived from each algorithm using the simulated health outcomes over 120 trajectories. The following results include the overall health outcomes for each policy, the action selection frequency that reflects on how the policy discovers the response type for each patient, and how accurately Update estimated the treatment effects.

3.4.1. Health outcomes

For both setups, the mean reward over 120 trajectories of the steady progression group is the highest among the three groups with the slow degrading group as the second highest (Table 3-6). This is not surprising as patient progression patterns define the transition among health states regardless of the policy performance.

Table 3-6. Mean health outcomes of each policy by progression groups for two setups

Setup	Policy	Fast degrading	Slow degrading	Steady
1	Human intuition	2.603	2.762	3.048
	Update	2.714	2.803	2.993
	FQI-Ridge	2.629	2.782	3.025
	FQI-RF	2.596	2.768	3.004
2	Human intuition	2.609	2.778	3.006
	Update	2.697	2.800	2.999
	FQI-Ridge	2.636	2.747	3.004
	FQI-RF	2.601	2.769	2.974

We performed one-way ANOVA to determine the difference in the performances of the four polices in each progression group. If the results show that the policies perform significantly different, we then performed a pairwise Tukey comparison test to identify whether our proposed policies outperform Human intuition. In Setup 1, we found that the health outcomes among the four policies are significantly different only in the fast-degrading group (Table 3-7). Within the fast-degrading group, Update provides the highest health outcomes and is the only policy that significantly outperforms Human intuition (Table 3-8). On the other hand, the two FQI approximators perform similarly to Human intuition.

In Setup 2, the health outcomes among the four policies are significantly different in the fast and slow degrading group (Table 3-7). Similar to setup 1, Update significantly outperforms other policies in the fast-degrading group, while there is no significant between Update and each of the two FQI approximators (Table 3-9). Within the slow-degrading group, all policies perform similarly except for Update which provides significantly higher health outcomes compared to FQI-Ridge.

We also investigated the average number of times that each health state appears in each patient trajectory by progression group for both setups (Figure 3-3). We observed that Update consistently brings a larger portion of health states to 'level 1' per patient on average. However, in the steady progression of Setup 2, all policies equally lead the patients toward the healthiest state.

Table 3-7. The p-values from one-way ANOVA on the health outcomes

Setup	Fast degrading	Slow degrading	Steady
1	1.9550E-20*	0.1513	0.1124
2	6.1447E-14*	0.0030*	0.4461

*p-values < 0.05: there is a significant difference among the mean of the health outcomes from the four policies.

Table 3-8. The mean difference and p-values from Tukey HSD test on the health outcomes for setup 1

Policy 1	Policy 2	Fast degrading	
		Mean difference	p-value
Human Intuition	Update	0.1108	0.001*
	FQI-Ridge	0.0251	0.2169
	FQI-RF	-0.0074	0.9
Update	FQI-Ridge	-0.0858	0.001*
	FQI-RF	-0.1182	0.001*
FQI-Ridge	FQI-RF	-0.0325	0.0612

*p-values < 0.05: there is a significant difference between the mean of the health outcomes from the two policies.

Table 3-9. The mean difference and p-values from Tukey HSD test on the health outcomes for setup 2

Policy 1	Policy 2	Fast degrading		Slow degrading	
		Mean difference	p-value	Mean difference	p-value
Human Intuition	Update	0.0881	0.001*	0.0216	0.443
	FQI-Ridge	0.0276	0.1408	-0.0318	0.1244
	FQI-RF	-0.0083	0.9	-0.0095	0.9
Update	FQI-Ridge	-0.0606	0.001*	-0.0534	0.0014*
	FQI-RF	-0.0964	0.001*	-0.0311	0.1393
FQI-Ridge	FQI-RF	-0.0359	0.0281*	0.0224	0.4114

*p-values < 0.05: there is a significant difference between the mean of the health outcomes from the two policies.

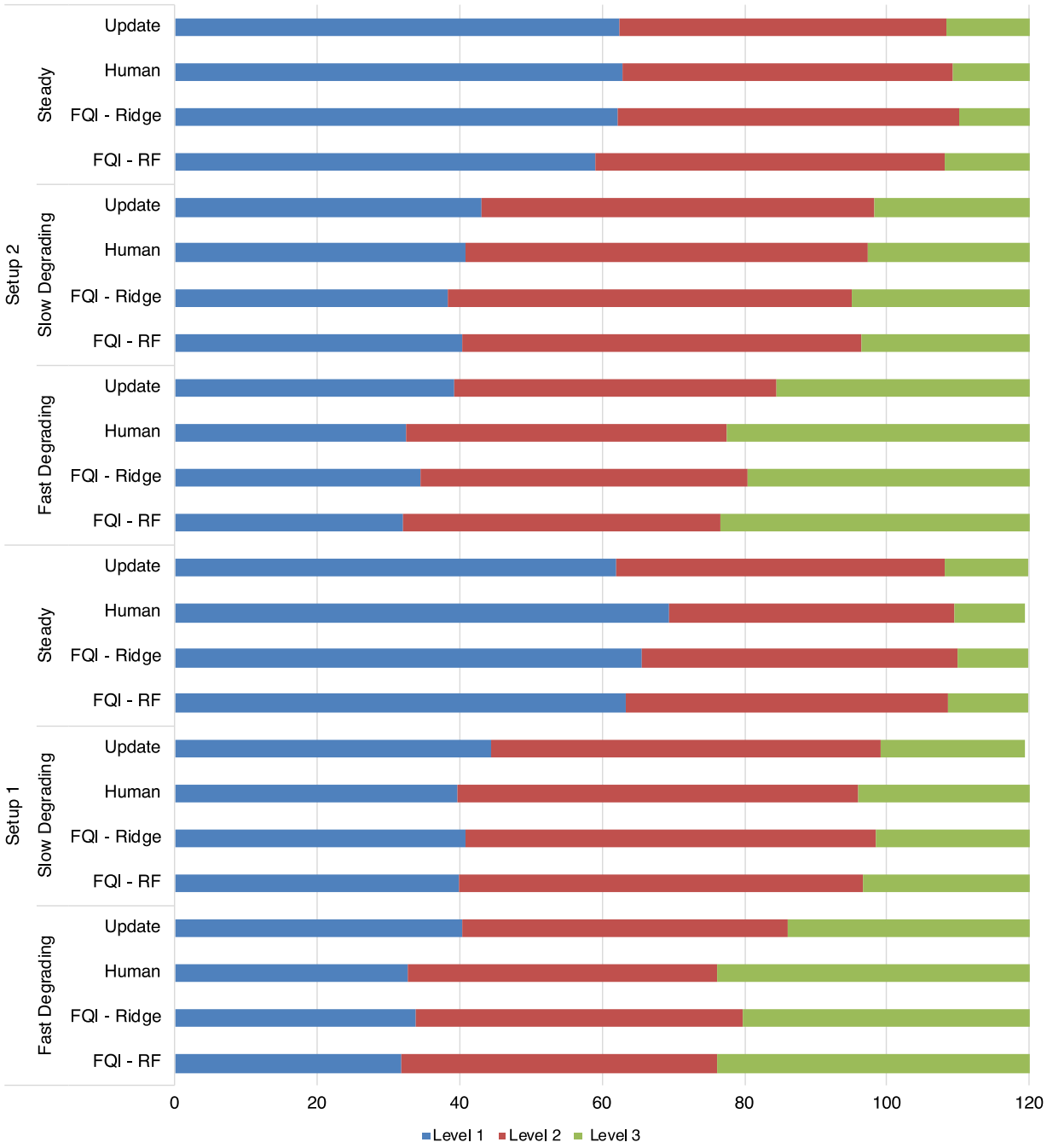


Figure 3-3. Average number of times each health state appears in one patient trajectory

3.4.2. Treatment frequency

To understand whether each policy can capture the patient's treatment response type, we aggregated the number of periods where each patient was on each treatment from all 120 patients and presented these numbers by treatment response types. Ideally, an optimal policy for our problem should only select the treatment that corresponds to a patient's response type.

Result showed that a mix between all three treatments were chosen by Human intuition with a bias towards the patient response type (Figure 3-4 and Figure 3-5). FQI-RF chooses all three treatments for all response types and its behavior is similar to Human intuition except for the steady group where it tends to choose treatment 2 most often in Setup 1. FQI-Ridge consistently places patients on a single treatment for each progression group regardless of the treatment response type – a behavior found in both setups. For fast degrading and steady progression groups, treatment 3, the most effective treatment on average, is chosen in both setups, while treatment 2 is chosen for slow degrading in Setup 2. Therefore, we can conclude that Human intuition and FQIs cannot capture heterogeneity in patients' treatment response type.

Update can provide the correct treatments to the right patients most of the time in all progression groups (Figure 3-6). When compared across progression groups, it can be observed that Update can capture heterogeneity most accurately in the fast-degrading progression. However in Setup 2, Update captures heterogeneity less accurately compared to Setup 1 as the policy selects a mix of three treatments more frequently in Setup 2.

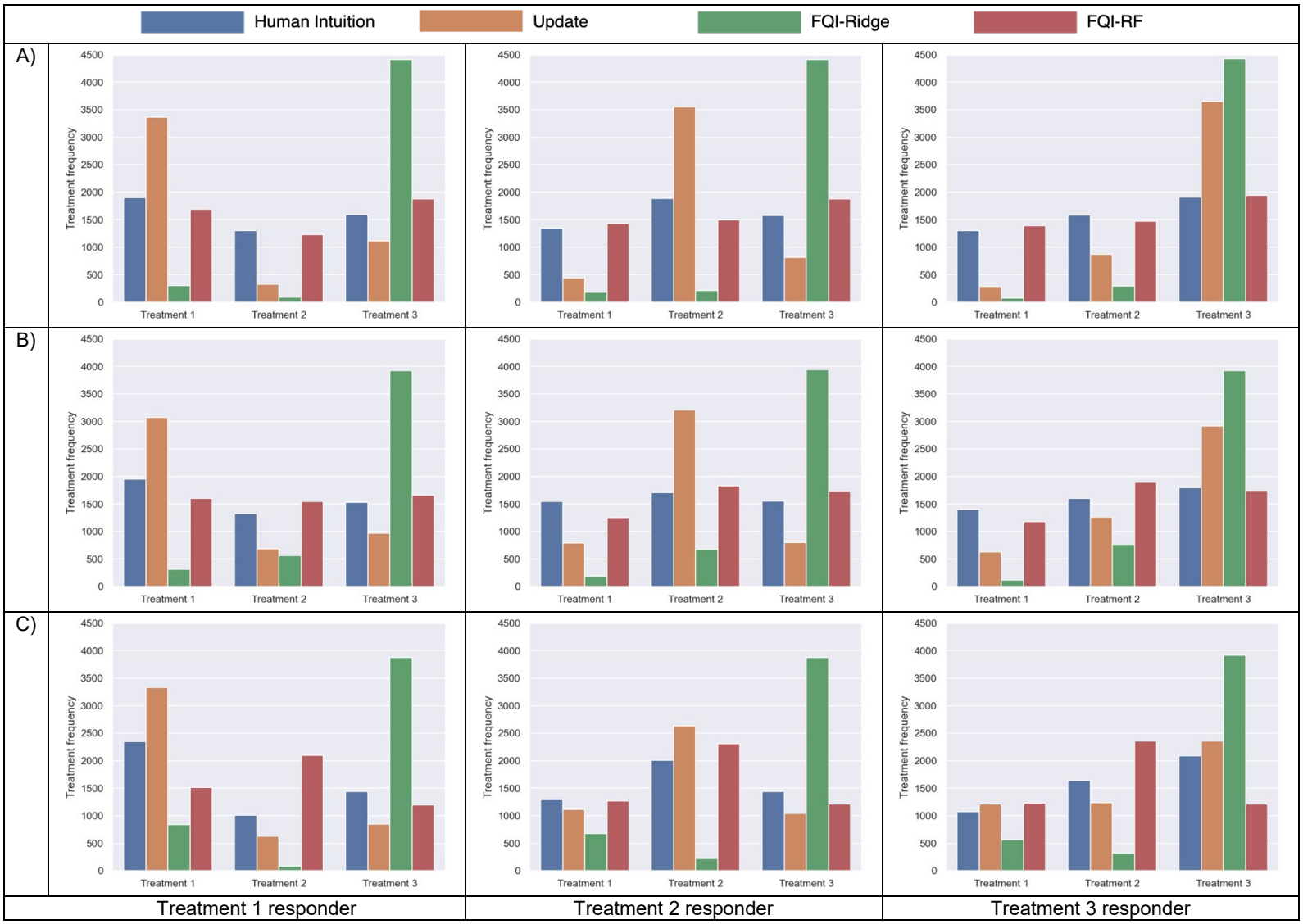


Figure 3-4. Treatment frequency in setup 1 compared across response types, with A) the fast-degrading group, B) the slow-degrading group, and C) the steady group

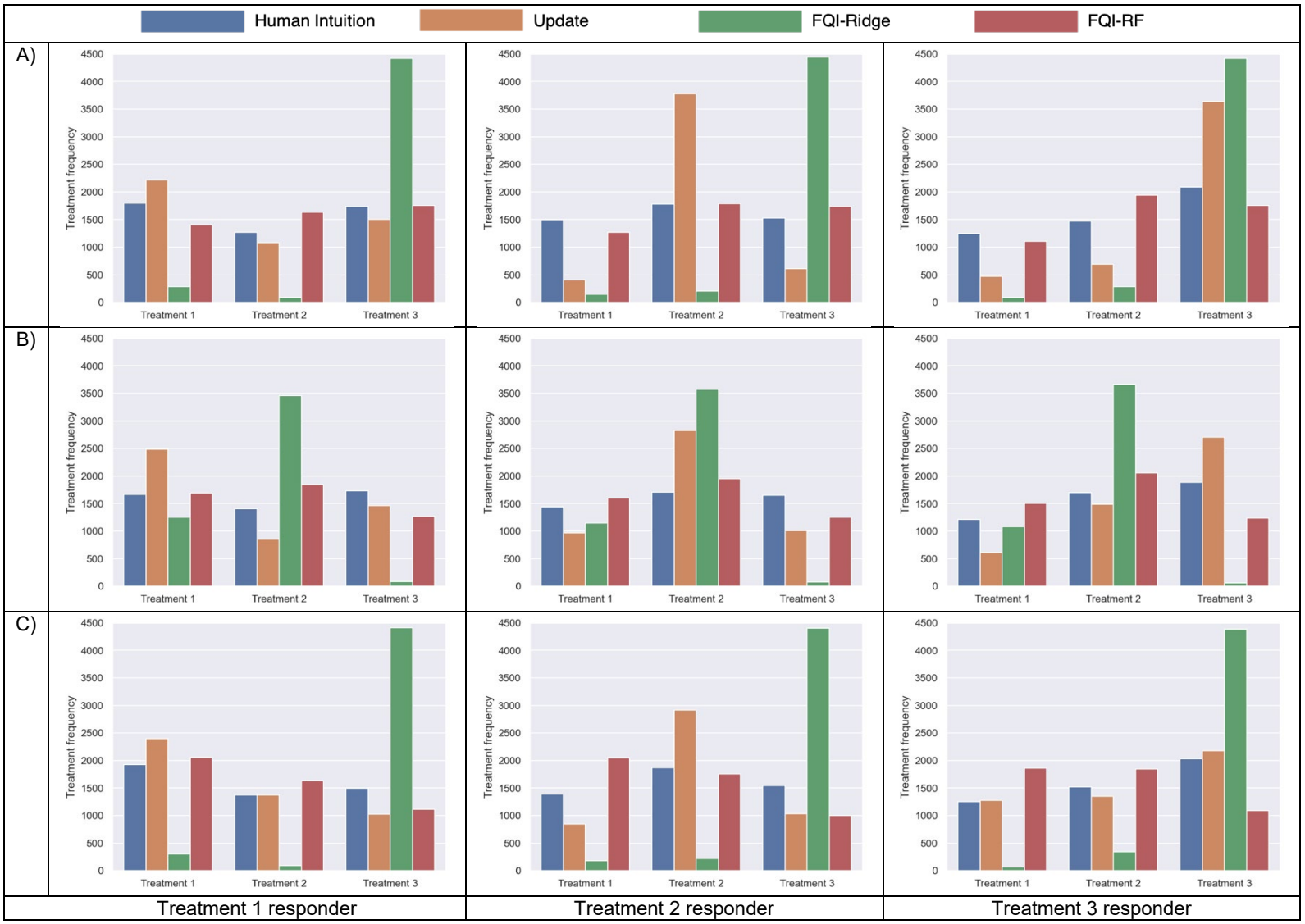


Figure 3-5. Treatment frequency in setup 2 compared across response types, with A) the fast-degrading group, B) the slow-degrading group, and C) the steady group

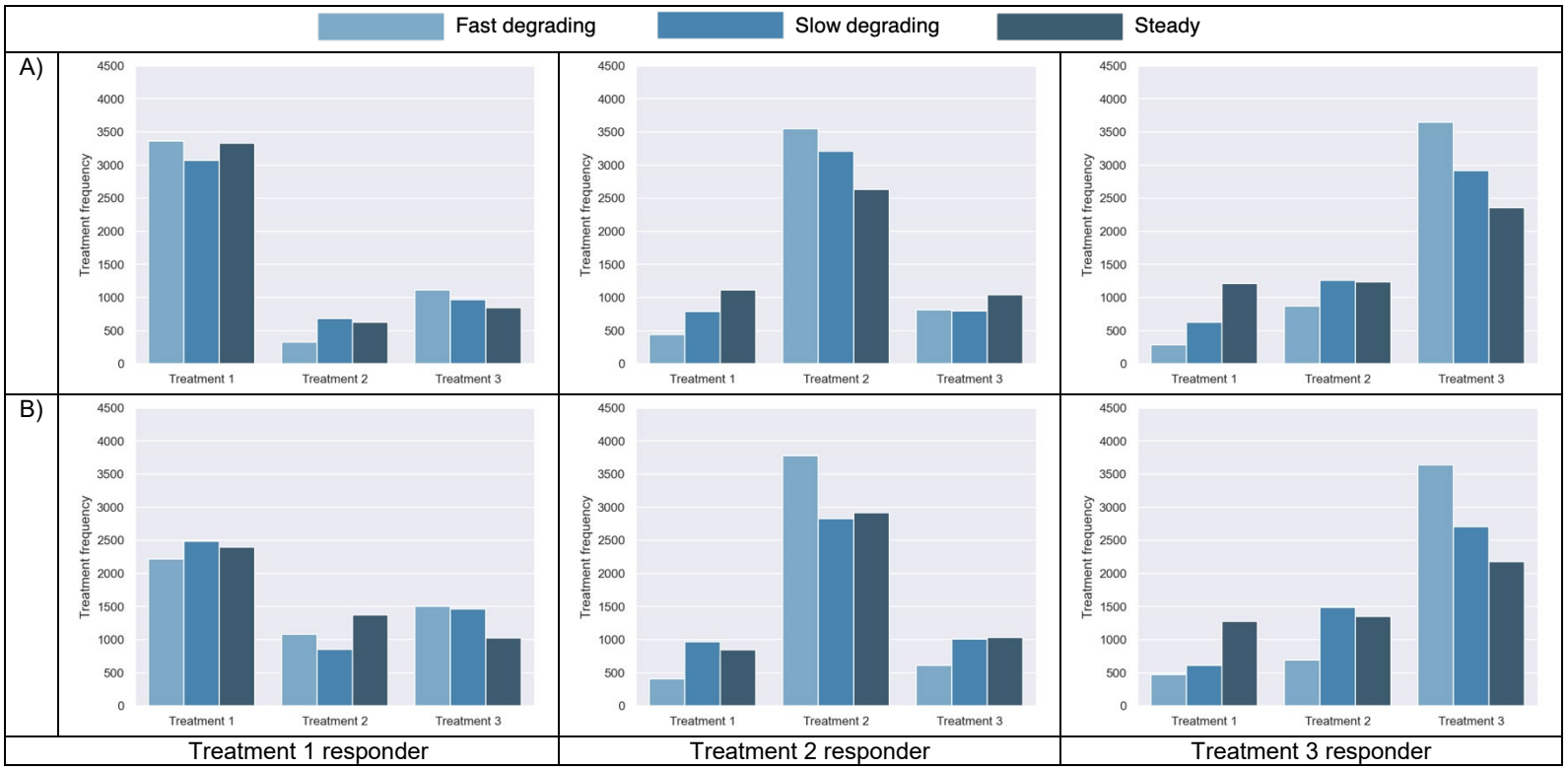


Figure 3-6. Treatment frequency of Update compared across response types and progression groups in A) setup 1 and B) setup 2

3.4.3. Treatment effect estimation

As Update directly estimates the treatment effects for each patient, we present the estimation errors for Setup 1 and Setup 2. Overall, Update can estimate treatment effects more accurately in fast degrading progression in both setups compared to other progressions. In both setups, the average mean squared errors (MSE) of all treatments in all progression groups are less than 0.1. However, the largest MSE can be as large as 0.6 (Figure 3-7). Note that these large values of MSE occur when MLE provides an estimated treatment effect of 1, which is the upper boundary of the treatment effects.

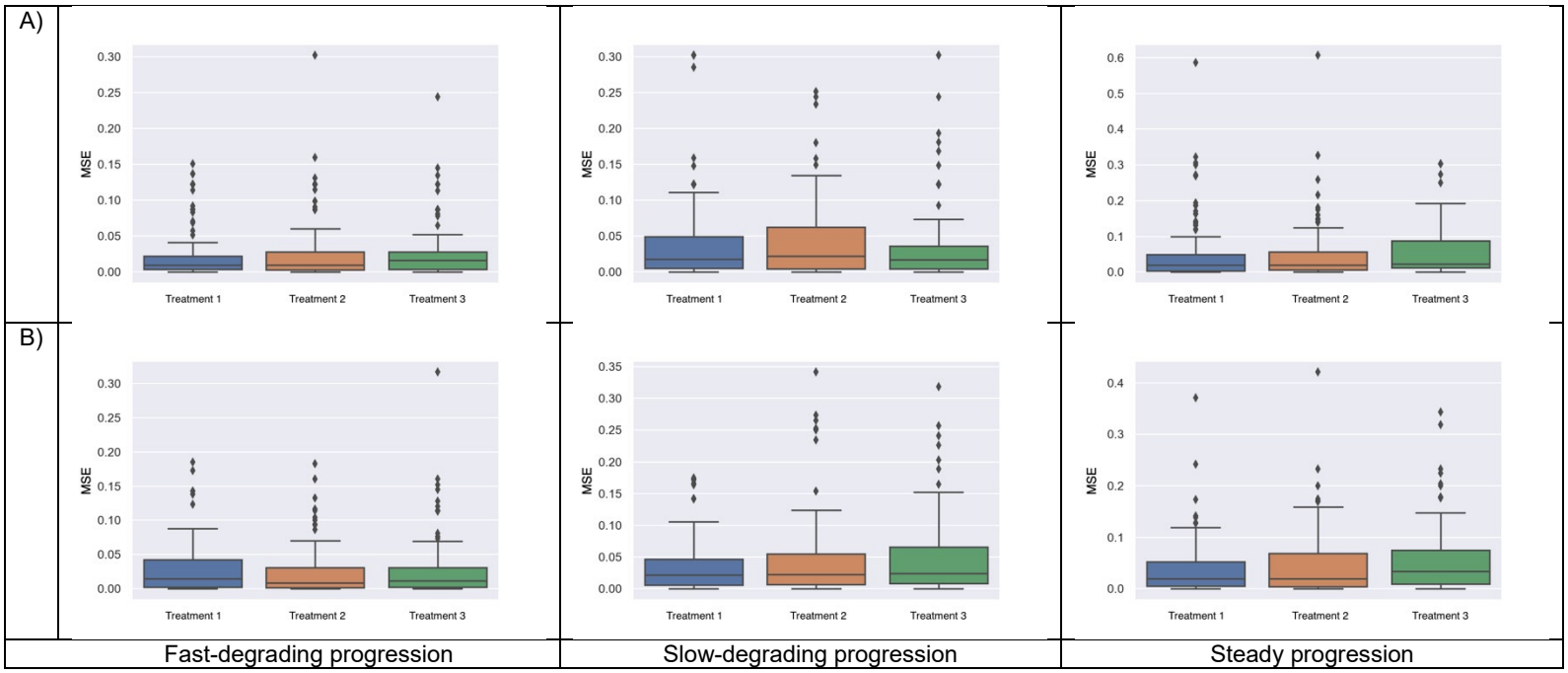


Figure 3-7. Mean squared error of treatment effect estimation by progression groups in A) setup 1 and B) setup 2

3.4.4. Sensitivity analyses

In the base-case results, we observed that Update can identify the patient response type correctly in both setups and significantly outperforms other policies in the fast-degrading group. To evaluate the robustness of Update, we explored its performance on five additional scenarios in each setup (Table 3-10).

According to one-way ANOVA and pairwise Tukey tests (Table 3-11 and Table 3-12), we observed that Update provides the highest health outcomes and outperforms other policies significantly in the fast-degrading group for both setups in all scenarios. However, there is no significant difference among the four policies in the steady progression group for all setups and scenarios. In the slow-degrading group, the overall performance of Update remains favorable where three scenarios of Setup 1 and one scenario of Setup 2 show that Update significantly outperforms at least one of the other policies.

In summary, the three major factors driving our results are 1) progression group, 2) heterogeneity level, and 3) parameter setting. Comparing across all progression groups, the more stable progression leads to similar performance among all policies in all setups and scenarios. With low heterogeneity level, Update outperforms others due to its relatively high accuracy in the treatment response type identification compared to the settings with high heterogeneity level. By varying several parameters (Table 3-10), we observed that increasing the health benefit gain from being in healthier states leads to an increase in the health outcomes of Update as its policy tends to bring patients toward healthier states the most. Additionally, dropout has no effect on our results

compared to the base-case results, while increasing treatment disutility in the more effective treatments leads to a decrease in the performance of FQI-RF.

Table 3-10. Scenario setting

Scenario	Description	Parameter setting
1	Being in healthier state incurs a gain in annual health utility of 0.30.	Utility = [0.80, 0.50, 0.20, 0]†
2	Being on antidepressant incurs 10 times side-effect burden compared to CBT	Treatment Disutility = [0.01, 0.1, 0.11]‡
3	Large dropout due to intolerance on antidepressant	Probability of dropout = [6.52%, 16.3%, 16.3%]‡
4	No difference in dropouts between CBT and antidepressant	Probability of dropout = [13.8%, 13.8%, 13.8%]‡
5	Large dropout due to intolerance and side-effect burden on antidepressant	A combination of Scenario 2 and 3; Treatment Disutility = [0.01, 0.1, 0.11]‡ Probability of dropout = [6.52%, 16.3%, 16.3%]‡

† The values of health utility correspond to the health states of mild, moderate, severe, and death, respectively. ‡ The values of treatment disutility/probability of dropout correspond to treatment 1 (CBT), treatment 2 (antidepressant), and treatment 3 (CBT + antidepressant), respectively.

Table 3-11. The p-values from one-way ANOVA on the health outcomes of setup 1 on each scenario

Scenario	Fast degrading	Slow degrading	Steady
1	3.0343E-15*	0.0079*	0.1158
2	1.2035E-09*	0.0142*	0.0550
3	9.6182E-06*	0.0314*	0.9468
4	1.8489E-16*	0.2011	0.3823
5	1.5982E-11*	0.0022*	0.2861

*p-values < 0.05: there is a significant difference among the mean of the health outcomes from the four policies.

Table 3-12. The p-values from one-way ANOVA on the health outcomes of setup 2 on each scenario

Scenario	Fast degrading	Slow degrading	Steady
1	3.7139E-11*	0.2459	0.2997
2	5.9485E-07*	0.0030*	0.0670
3	3.3800E-03*	0.1857	0.0909
4	2.7650E-12*	0.3220	0.8220
5	1.7532E-04*	0.1104	0.5343

*p-values < 0.05: there is a significant difference among the mean of the health outcomes from the four policies.

3.5. Discussion

We presented a treatment selection problem for patients with chronic depression with the objective of finding the best policy out of a rule-based algorithm (Human Intuition), a model-based algorithm (Update), and a model-free algorithm (FQI). The policy performances and robustness were evaluated on several simulation experimental setups specified by patient disease progression and heterogeneity in treatment effects. Results showed the Update algorithm provides the best policy which captures the heterogeneity in treatment response types and is robust towards changes in parameter setting. FQI-Ridge tends to select a single treatment for all response types and progression groups, while FQI-RF selects a mix of treatments with a bias towards the more effective treatment.

At first glance, Human Intuition performed relatively well on health outcomes compared to other algorithms according to our results. However, Human Intuition treats all patients the same regardless of their response types by its nature of being one-size-fit-all algorithm. There are various scenarios where providing suboptimal treatments could harm a patient's quality of life. Several examples of such scenarios that required treatment tailored to each patient are 1) when a patient's health can deteriorate quickly, 2) high risk of death, and 3) treatments have severe side effects.

Our study relies on the assumption that each patient responds best to one of the available treatments, which is quantified by treatment effect. Under this assumption, Update can identify the patient treatment response types most accurately among our algorithms due to its ability to estimate treatment effects by leveraging the knowledge of the disease progression and the effect of treatment effects on disease progression. FQI learned a policy by creating a model from the relationship between a patient's state, which composed of past disease severity levels and treatments, and the corresponding treatments. We believe that FQI cannot accurately identify treatment response types because its model cannot directly observe treatment effect but can only focus on the degree of importance of each component in a patient's state.

Several studies in mental healthcare have demonstrated success in applying reinforcement learning to their treatment selection problem [64,75,77]. However, our results seem to contradict with those findings. One of the reasons could be that we did not perform feature selection when developing our model approximator for Q-function, and therefore, we utilize all information collected from our holding period. This may have led to a sparse feature vector or model overfitting. Additionally, reinforcement learning relies heavily on a reward function as it specifies how an agent (decision maker) select an action. In this study, we have only considered a general reward function of health outcomes consisting of health utility and treatment disutility, which are widely used in health economics. As such, there might be other reward functions that lead to a better model approximator for the Q-function in our problem setting, which is beyond the scope of this chapter.

We observed that patients' treatment selections are influenced by their disease progression patterns. Using simulated patient trajectories, we have shown that steady progression leads to similar health outcomes among all treatment policies regardless of the correctness in selecting treatment, while the model-based policy clearly outperforms others among patients with the fast-degrading progression. This is due to patients with steady progression, once transitioned into a better health state will tend to remain in that state regardless of treatment. As a result, our work

highlights the necessity in clustering patients according to their distinct trajectory patterns as one of the crucial steps in personalized treatment. Similar to our findings, Lin et al. has shown that clustering patients with depression by their trajectory pattern leads to better monitoring strategies [89].

If we view Human Intuition as a clinician following clinical guidelines, our proposed algorithms or data-driven artificial intelligence methods could be considered as decision support tools that accumulate data from interactions between human clinicians and patients, detect hidden patterns, and suggest decisions that may lead to the best outcomes [90]. As various sources of clinical data become accessible, a clinician would have an overwhelming amount of data to process in decision making, and therefore, artificial intelligence tools would be increasingly needed as support tools [91]. Although there are concerns in mental healthcare that machine-suggested decisions might override a clinical expert's judgement, the suggested decisions would be delivered to human clinicians and patients, who make the final judgement [90].

To apply our model-based algorithm in practice, there are three necessary tasks that need to be completed: 1) categorize patients into subgroups by trajectory pattern, 2) estimate a subgroup disease progression matrix or non-treatment matrix, and 3) find the imposed structure of treatment matrix with treatment effect incorporation. The first task can be completed in several ways depending on data availability. Without data available, an observational study on patients with no treatment needs to be conducted. Once the study has been conducted or data is available, we would have patient trajectories with no treatment. Then, we can simply use a clustering technique to find their distinct patterns. Some examples of clustering techniques are K-means clustering [92], latent class growth analysis [92,93], growth mixture modeling [92,94], or collaborative learning method [95]. To complete the second task, we can estimate a subgroup non-treatment matrix given the patient trajectories of disease symptoms from the first task using the maximum likelihood estimation, a common approach for transition matrix estimation [96]. Lastly, to understand how treatment effect is incorporated to the non-treatment matrix, a clinical trial study needs to be conducted to learn both the treatment effectiveness and the imposed structure of treatment progression matrix.

Our approaches learn their models and apply them to patients iteratively. This iterative learning process is similar to how a clinician meets patients and is considered offline learning. In each visit, a clinician would observe how patients respond to treatment after a fixed period and make a new decision based on those responses and patient history. Similarly, we leverage accumulated data from all previous iterations to train a new model, and therefore, the model provides an optimal solution with the available data. The model is then used in decision making for the specified period. Hypothetically, our learning process is similar to an adaptive group sequential trial where treatment modification is allowed at predefined interim points.

Furthermore, our framework can be extended to treatment selection problems with more than three treatment types by simply expanding the action space M to the number of available treatments. With more treatment types, a multi-arm trial should be conducted to learn a population treatment effect $\bar{\rho}^m$. To compensate for unselected treatments due to large action space, an exploration may be allowed in our algorithms. Moreover, we believe our work can be generalized to personalized treatment planning for other chronic diseases that fall under the following assumptions: 1) a

disease progression follows a Markov model, 2) a patient' observations are fully observable and monitored regularly, 3) a patient requires a long-term adaptive treatment, and 4) the treatment space is discrete. In addition, our work could be considered as a general framework for an emerging research topic in healthcare, just-in-time adaptive intervention, which generally aims to change patient behavior to prevent negative health outcomes using individual collected data through sensing devices [97]. Some examples of such studies are adaptive stress intervention [98] and physical activity suggestion [99]. As an example of adapting our algorithm for physical activity suggestion, Update can estimate the treatment effect for each choice of activities with MLE based on daily collected data and provide individual suggestion at each decision time point based on the expected reward and individual treatment transition.

This study has several limitations. First, we had limited access to real patient trajectory data and therefore applied our methods to simulated patient trajectories using parameters found from depression-related studies. Second, we assumed all patient observations were monitored at every period. However, due to the increasing use of digital tools and virtual care for mental health, patient observations can now be observed regularly via digital assessment or online clinics. Several studies have found that digital mental health care is as effective as in-person care [100–102]. Third, the accuracy of treatment effect estimation relies on sufficient observations from a holding period. In practice, a duration of the holding period could be shorter than that of our numerical experiment leading to a small number of observations. To address this issue, we could estimate the treatment effect of a treatment using concatenated observations of a patient being on that treatment instead of using only the observations from the holding period.

In conclusion, we presented a treatment selection problem and formulated our problem as an MDP. Using simulated patient trajectories, we evaluated the performance of policies derived from rule-based algorithms, model-based algorithms, and model-free algorithms. Given the knowledge of disease progression prior to treatment and the incorporation of treatment effects, the model-based algorithm has an advantage over the model-free algorithm and can successfully capture the heterogeneity in patient response type. Our work provides a step towards the design of personalized treatment strategy for patients with chronic diseases.

4. Estimating patient preference in clinical decision making using ranked trajectories

4.1. Introduction

With emerging technologies in artificial intelligence, big data, and sensor systems for healthcare, the concept of tailoring medical treatment towards individual characteristics, or personalized medicine, has been widely applied in clinical decision making [91,103,104]. In some cases, these treatments should not only be tailored to the individual's characteristics but also their preferences, as patients' priorities may include the trade-off between health benefits and their personal risks [2]. By incorporating preferences into clinical decisions, patients tend to have increased satisfaction with their healthcare provider, are more willing to accept and adhere to their preferred treatment [2,3], and could help improve their overall health outcomes [105].

In clinical decision making, various preference elicitation methods are used to quantify patient preferences, such as matching methods, discrete choice experiments, and multi-criteria decision analysis [106]. A detailed explanation of the three methods can be found in [107] and [106]. Among these three methods, discrete choice experiments are most often used in clinical decision making, specifically: direct choice, rating and ranking methods [106]. A complete ranking exercise over all alternatives is an elicitation technique that provides information for all pair-wise comparisons of all alternatives [107]. With a small number of alternatives, ranking is among the simple methods that requires low cognitive effort [106].

Generally, personalized treatment problems are formulated as sequential decision-making problems under uncertainty, which is commonly modelled as a Markov decision process (MDP). MDP problems are often solved for their optimal policies using reward functions. When taking into account preferences, the reward function now has multiple competing objectives. Several studies [108–111] modelled their reward functions by parametrizing with an individual's preference, which in most settings, are assumed to be unknown.

In a treatment selection problem for schizophrenia patients, Lizotte et al. [108] modelled their reward function as a linear combination of preferences over multiple objectives. They proposed a reinforcement learning algorithm, specifically Fitted Q-iteration where the Q-function is linear-piecewise function in preferences, and solved for optimal actions for all possible preference values and all states. In a similar application, Butler et al. [109] also considered a linear reward function in preferences and combined an item response model and Q-learning to simultaneously estimate patient preferences using inputs from questionnaires to provide optimal treatment plans. In a liver transplantation problem [111], inverse optimization techniques were applied to estimate patient's preferences over health states while assuming patients act optimally over their unknown health state valuations.

Outside of literature in healthcare, there exists research involving MDP problems with unknown reward functions with the objective of finding an optimal policy. One of the most popular techniques to solve for such problem is inverse reinforcement learning which is used to recover the unknown reward function [112–115]. Abbeel and Ng [113] proposed an apprenticeship learning with the goal of learning a policy whose performance is close to the expert's policy, which

is unknown. They assumed that they have access to the expert's trajectories. Their unknown reward function is parametrized by an unknown weight vector that is assumed to be linear. Considering a similar problem to Abbeel and Ng, Ikenaga and Arai [116] proposed an algorithm based on an apprenticeship learning to estimate the weight vector given the expert's trajectories, which is assumed to be based on an optimal policy.

However, the expert's trajectories may not necessarily come from an optimal policy, and we may not want to imitate the expert's policy. Therefore, some studies have relaxed this assumption such as studies on learning from human demonstration that includes failed trajectories [117,118]. Brown et al. [119] also assumed that the existing demonstration (trajectories) are suboptimal. Their study aims to estimate the unknown reward function and find a policy that outperforms the suboptimal trajectories, given access to the ranking over the trajectories.

In this chapter, we are interested in learning an individual's preference over two competing objectives and finding an optimal policy for a sequential decision-making problem formulated as a Markov decision process when the reward function is parameterized by an individual's preference. We assume we have access to an expert's trajectories that are suboptimal for an individual's true reward, a complete ranking of trajectories provided by an individual, and the reward function is a linear combination in preference. The assumption of a linear reward function is common in problems with multiple objectives [109] and in preference elicitation [120]. In this work, we introduce new algorithms that can estimate preferences by iteratively learning from multiple sets of complete rankings of suboptimal trajectories and provide an optimal policy based on the recovered reward function using the learned preference. By utilizing information from complete ranking, which is a simple preference elicitation, our algorithms provide an accurate estimation of preference without relying on questionnaire models, leading to policies that outperform the initial suboptimal policy.

This chapter is organized as follows. Section 4.2 introduces the methodologies used in formulating our problem as an MDP, the trajectory rankings, and our algorithms. In Section 4.3, we describe the experimental setup for our simulation and present numerical results on our algorithms' performance at baseline along with one-way sensitivity analysis to determine relative performance. Lastly, we conclude with a discussion of our results, methods, and limitations in Section 4.4.

4.2. Methodology

4.2.1. Problem formulation

We consider a problem of learning an individual's preference and finding an optimal policy in a Markov decision process (MDP) with finite states and time horizon given an expert's demonstration or trajectories, which comes from an unknown policy, along with complete rankings of multiple trajectory sets provided by the individual. We assume each set of trajectories is ranked according to scores calculated from an individual's true reward function, which is a weighted summation of two objectives such as benefit gained from being in one health state and side effect burden from using a treatment. We assume an individual has their own preferences towards the two objectives, however, it is difficult to be quantified as an exact number. Therefore, we instead ask an individual to rank a set of trajectories and use the rankings to learn an individual's

preference. An individual's true reward function $R(\boldsymbol{w}_p, s)$ is parametrized by an unknown individual preference $\boldsymbol{w}_p$ and observable individual state. As the true individual preference is unknown to decision makers, we formulate our problem as an MDP with an unknown reward function using the following definitions:

States: Let S denote a finite discrete state space and s_t denote a state an individual is in at a time period t where $s_t \in S$.

Actions: Let A denote a finite discrete action space and a_t denote an action taken at a time period t where $a_t \in A$.

Transition probability: Let $p(s_{t+1}|s_t, a_t)$ denote a probability of transitioning from state s_t to state s_{t+1} when the action a_t is chosen at a time period t. We assume that the transition probability is known.

Preference: Let $\boldsymbol{w}$ denote a preference weight vector where $\boldsymbol{w} \in [0,1]^K$. Let w^k denote the k[th] element of the vector $\boldsymbol{w}$ where $\sum_{k=1}^{K} w^k = 1$ and K is the dimension of the vector $\boldsymbol{w}$.

Reward function: For any $\boldsymbol{w}$, the reward function $R(\boldsymbol{w}, s)$ is a linear combination of two known functions. Let $\boldsymbol{r}(\boldsymbol{s})$ denote a function vector where $\boldsymbol{r}(\boldsymbol{s}) \in [0,1]^K$ and $r^k(s)$ denote the k[th] element of the vector $\boldsymbol{r}(\boldsymbol{s})$. The reward function can be expressed as follows:

$$R(\boldsymbol{w}, s) = \boldsymbol{w^T r}(\boldsymbol{s})$$

For $K = 2$, we have the following reward function:

$$\begin{aligned} R(\boldsymbol{w}, s) &= w^1 \times r^1(s) + w^2 \times r^2(s) \\ &= w^1 \times r^1(s) + (1 - w^1) \times r^2(s) \end{aligned}$$

Policy: A policy π is a mapping from states to actions for each time period t where $t = 1, 2, \dots, T$. Let π_{w_e} denote the unknown expert policy, which is optimal based on $R(\boldsymbol{w_e}, s)$ where w_e is the expert's preference and is unknown to the decision maker. For any $\boldsymbol{w}$, an optimal policy π_w^* solved under MDP is defined as follows:

$$\pi_w^* = \underset{\pi}{\operatorname{argmax}}\, E\left[\sum_{t=0}^{T} \gamma^t R(\boldsymbol{w}, s_t)|\pi\right]$$

Trajectory: Let a trajectory τ denote a sequence of states for all time periods, which is expressed as $\tau = (s_0, s_1, \dots s_T)$.

Feature expectation: Similar to Abbeel and Ng [113], we define a vector $\boldsymbol{\mu}(\pi)$ as the expected discounted summation of rewards gained over a time horizon T given a policy π. Let $\mu^k(\pi)$ denote the k[th] element of the vector $\boldsymbol{\mu}(\pi)$ which can be expressed as follows:

$$\mu^k(\pi) = E\left[\sum_{t=0}^{T} \gamma^t r^k(s_t)|s_0 = s, \pi\right]$$

where $\boldsymbol{\mu}(\pi) \in \mathbb{R}^k$ and γ is a discount factor.

Given a set $\mathcal{T}$ of M trajectories generated from a policy π where $\mathcal{T} = \{\tau_i\}_{i=1}^{M}$ and $\tau_i = \left(s_0^{(i)}, s_1^{(i)}, \dots s_T^{(i)}\right)$, we can calculate the empirical feature expectation $\hat{\mu}^k(\pi)$ as follows:

$$\hat{\mu}^k(\pi) = \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\gamma^t r^k(s_t^{(i)})$$

Value function: A value function $V_\pi(\boldsymbol{w}, s)$ is the expected reward under policy π when starting in state s. Given a probability distribution $\mathcal{D}$ of an initial state s_0, $V_\pi(\boldsymbol{w}, s)$ for can be expressed as follows:

$$\begin{aligned} V_\pi(\boldsymbol{w}, s) &= E\left[\sum_{t=0}^{T}\gamma^t R(\boldsymbol{w}, s_t)|s_0 = s, \pi\right] \\ &= \boldsymbol{w^T} E\left[\sum_{t=0}^{T}\gamma^t \boldsymbol{r(s_t)}|s_0 = s, \pi\right] \end{aligned}$$

Using the definition of the feature expectation, we can express $V_\pi(\boldsymbol{w}, s)$ in terms of $\boldsymbol{\mu}(\pi)$ as follows:

$$V_\pi(\boldsymbol{w}, s) = \boldsymbol{w^T}\,\boldsymbol{\mu}(\pi)$$

Let $V_{\pi_w^*}(\boldsymbol{w}, s)$ denote an optimal value function when the reward function is parametrized by $\boldsymbol{w}$ and is expressed as follows:

$$V_{\pi_w^*}(\boldsymbol{w}, s) = \max_\pi V_\pi(\boldsymbol{w}, s)$$

As $R(\boldsymbol{w}, s_t)$ is a linear function in $\boldsymbol{w}$, we can prove that $V_{\pi_w^*}(\boldsymbol{w}, s)$ is a piecewise linear function in $\boldsymbol{w}$ by showing that $V_{\pi_w^*}(\boldsymbol{w}, s)$ follows Definition 1 as shown in Proposition 1.

Definition 1: A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is piecewise-linear if there exist $c_1, \dots, c_N \in \mathbb{R}^n$ and $d_1, \dots, d_N \in \mathbb{R}$ s.t. $f(x) = \max_{j\in[N]} \boldsymbol{c_j^T x} + d_j$

Proposition 1: $V_{\pi_w^*}(\boldsymbol{w}, s)$ is a piecewise linear function in $\boldsymbol{w}$.

Proof.

Let $\mathcal{W}$ denote a finite preference weight space where $\mathcal{W} = \{\boldsymbol{w_j}\}_{j=1}^{N}$.

Let π_j^* denote an optimal policy solved under the MDP with a reward function $R(\boldsymbol{w_j}, s)$.

Let π_w^* denote an optimal policy solved under the MDP with a reward function $R(\boldsymbol{w}, s)$ for any $\boldsymbol{w} \in \mathcal{W}$.

Let $\prod$ denote a set of N candidate policies where $\prod = \{\pi_j^*\}_{j=1}^{N}$

For $N = 2$, we have two optimal policies π_1^* and π_2^* corresponding to $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ respectively.

When $\boldsymbol{w} = \boldsymbol{w_1}$, we have

$$V_{\pi_1^*}(\boldsymbol{w_1}, s) = \max_{\pi\in\prod} \boldsymbol{w_1}^T\,\boldsymbol{\mu}(\pi) = \max\{\boldsymbol{w_1}^T\,\boldsymbol{\mu}(\pi_1^*), \boldsymbol{w_1}^T\,\boldsymbol{\mu}(\pi_2^*)\} = \boldsymbol{w_1}^T\,\boldsymbol{\mu}(\pi_1^*)$$

When $\boldsymbol{w} = \boldsymbol{w_2}$, we have

$$V_{\pi_2^*}(\boldsymbol{w_2}, s) = \max_{\pi \in \Pi} \boldsymbol{w_2}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi) = \max\{\boldsymbol{w_2}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_1^*), \boldsymbol{w_2}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_2^*)\} = \boldsymbol{w_2}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_2^*)$$

Therefore, we can write $V_{\pi_1^*}$ as a function of $\boldsymbol{w}$ as follows:

$$\begin{aligned} V_{\pi_w^*}(\boldsymbol{w}, s) &= \begin{cases} \boldsymbol{w_1}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_1^*) \\ \boldsymbol{w_2}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_2^*) \end{cases} \\ &= \begin{cases} \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_1^*), & \boldsymbol{w} = \boldsymbol{w_1} \\ \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_2^*), & \boldsymbol{w} = \boldsymbol{w_2} \end{cases} \\ &= \max\{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_1^*), \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_2^*)\}, \quad \forall \boldsymbol{w} \in \{\boldsymbol{w_1}, \boldsymbol{w_2}\} \end{aligned}$$

For any N, we can express $V_{\pi_w^*}(\boldsymbol{w}, s)$ as a function of $\boldsymbol{w}$ as follows:

$$\begin{aligned} V_{\pi_w^*}(\boldsymbol{w}, s) &= \max_{j \in [N]} \{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_1^*), \dots, \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_j^*), \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_{j+1}^*), \dots, \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_N^*)\}, \quad \forall \boldsymbol{w} \in \mathcal{W} \\ &= \max_{j \in [N]} \{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\mu}(\pi_j)\} = \max_{j \in [N]} \{\boldsymbol{\mu}(\pi_j)^{\boldsymbol{T}} \boldsymbol{w}\}, \quad \forall \boldsymbol{w} \in \mathcal{W} \end{aligned}$$

By Definition 1, $V_{\pi_w^*}(\boldsymbol{w}, s)$ is a piecewise-linear function in $\boldsymbol{w}$ with $c_j = \boldsymbol{\mu}(\pi_j)$ and $d_j = 0$.

In Figure 4-1, we show a numerical example of Proposition 1 when $K = 2$ and $N = 11$. This example considers an MDP with 12 states, where each state consists of health state {1, 2, 3, D} with D as an absorbing state and treatment state {1, 2, 3}, and 3 actions of {1, 2, 3}. Each action has a transition probability matrix defined in Table 4-1. A reward function is parametrized by preference weight $\boldsymbol{w}$ and is a linear combination of a reward function over health state and a reward function over treatment (Table 4-2). When a health state is D, a reward function takes a value of 0. The objective of the MDP is to find an optimal policy that maximizes an expected reward over a finite time horizon T of 200 periods. We denote $\boldsymbol{w}$ as $[w, 1-w]^T$ and define the preference weight space from 0.0 to 1.0 with 0.1 increments. Each colored line represents $\boldsymbol{w}^T \hat{\boldsymbol{\mu}}(\pi_j^*)$ for all values of w where π_j^* is optimal for the reward $R(\boldsymbol{w_j}, s)$ and $\pi_j^* = \underset{\pi}{\operatorname{argmax}} E\left[\sum_{t=0}^{T} \gamma^t R(\boldsymbol{w_j}, s_t) | \pi\right]$. We calculated $\hat{\mu}(\pi_j^*)$ using 5000 trajectories (M = 5000). For example, the policy values of "μ @ $w = 0.1$" is calculated using π_j^*, which is optimal for $\boldsymbol{w_j} = [0.1, 0.9]^T$. The black line is the optimal policy values for all preference weights, which shows that the proof is consistent with the numerical example.

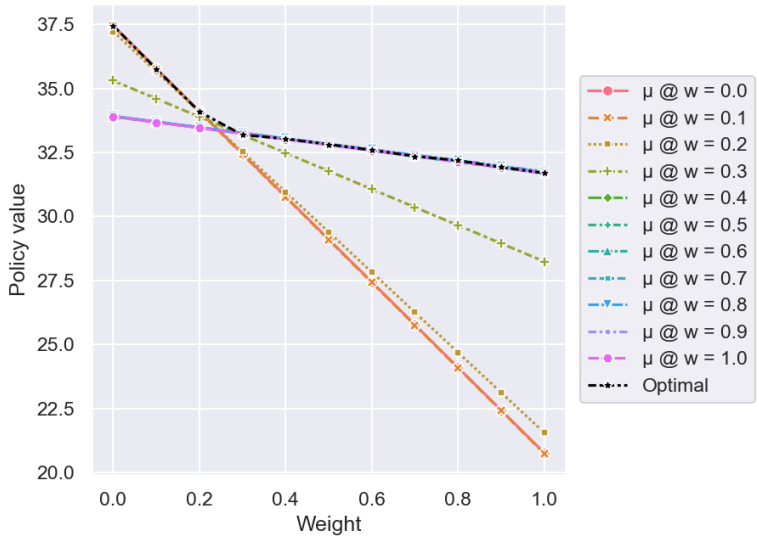


Figure 4-1. Policy values varied by preference weights where each policy is optimal for its corresponding preference weight (colored lines). The optimal policy values varied by preference weight are shown as a piecewise-linear function (black line).

4.2.2. Trajectory ranking

Given a set of trajectories $\{\tau_i\}_{i=1}^{M}$ generated by the same unknown policy π, we can calculate the empirical feature expectation for each trajectory i as follows:

$$\hat{\mu}_{(i)}^{k}(\pi) = \sum_{t=1}^{T} \gamma^t r_k\left(s_t^{(i)}\right)$$

Let $sc\left(\boldsymbol{w}, \hat{\boldsymbol{\mu}}(\pi)\right)$ be the score function of a trajectory where $sc\left(\boldsymbol{w}, \hat{\boldsymbol{\mu}}(\pi)\right) = \boldsymbol{w}^T \hat{\boldsymbol{\mu}}(\pi)$ and w is unknown. Note that $sc\left(\boldsymbol{w}, \hat{\boldsymbol{\mu}}(\pi)\right)$ is similar to the policy value function. The only difference is that $\hat{\boldsymbol{\mu}}(\pi)$ is estimated from a single trajectory.

Let $\tau_{[i-1]}$ denote the $(i-1)^{\text{th}}$ ranked trajectory and $\tau_{[i]}$ denote the i^{th} ranked trajectory.
Let $\tau_{[i-1]} \succcurlyeq \tau_{[i]}$ represent that $\tau_{[i-1]}$ is more preferable than $\tau_{[i]}$ and therefore $sc_{[i-1]} \geq sc_{[i]}$.
Let $W_{[i]}$ denote a set of feasible preference weights that provide $sc_{[i]}$ such that $\tau_{[i]}$ has the i^{th} rank.

Using Proposition 1, we can find the feasible preference weights by expressing the score function as a piecewise-linear function in $\boldsymbol{w}$ considering a ranked trajectory set if the preference weight space is finite and all possible weights in the space are known.

Considering a ranked trajectory set $\left\{\tau_{[i]}\right\}_{i=1}^{M}$, we can find the feasible weights $W_{[1]}$ such that $\tau_{[1]}$ has the highest score:

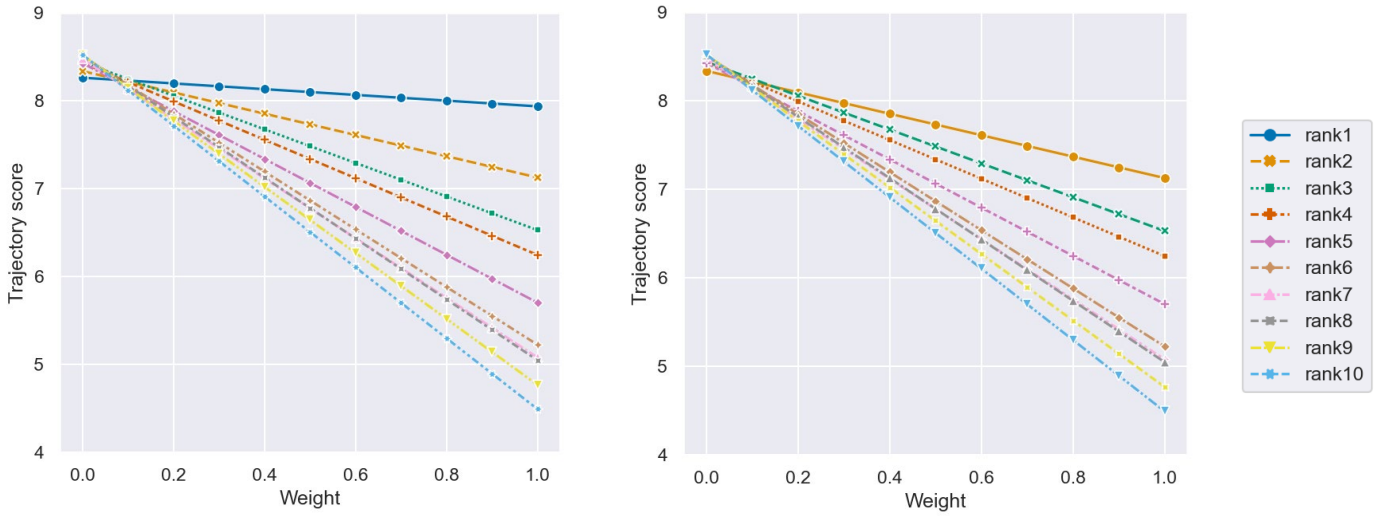
$$sc_{[1]}\left(\boldsymbol{w}, \hat{\boldsymbol{\mu}}_{[1]}(\pi)\right) = \max_{i \in [M]} \left\{\boldsymbol{w}^T \hat{\boldsymbol{\mu}}_{[i]}(\pi)\right\}$$

Considering a ranked trajectory set $\{\tau_{[i]}\}_{i=2}^{M}$, we can find the feasible weights $W_{[2]}$ such that $\tau_{[2]}$ has the highest score:

$$sc_{[2]}\left(\boldsymbol{w}, \hat{\boldsymbol{\mu}}_{[2]}(\pi)\right) = \max_{i\in[2,\dots,M]} \{\boldsymbol{w}^T \hat{\boldsymbol{\mu}}_{[i]}(\pi)\}$$

By searching for the feasible preference weights using a full set $\{\tau_{[i]}\}_{i=1}^{M}$ to the smallest set $\{\tau_{[i]}\}_{i=M-1}^{M}$, we can get the feasible preference weights providing scores that could match all rankings as $W = \cap_{i=1}^{M} W_{[i]}$.

In Figure 4-2, we show a numerical example of feasible preference weights for$\{\tau_{[i]}\}_{i=1}^{M}$ and $\{\tau_{[i]}\}_{i=2}^{M}$ when $K = 2$, $M = 10$, and $T = 12$. We denote $\boldsymbol{w}$ as $[w, 1-w]^T$and define the preference weight space as the values from 0.0 to 1.0 with 0.1 increments. Each line represents $\boldsymbol{w}^T \hat{\boldsymbol{\mu}}_{[i]}(\pi)$ for all values of w where $\hat{\boldsymbol{\mu}}_{[i]}(\pi)$ is estimated from trajectory $\tau_{[i]}$. For example, the policy values of "$rank$ 1" is calculated using the 1st ranked trajectory. In Figure 4-2 (A), the scores of the rank 1 trajectory shown in blue are the highest when the preference weight w is larger than 0.1. In Figure 4-2 (B), the scores of the rank 2 trajectory shown in orange are the highest when the preference weight w is larger than 0.2. Therefore, the feasible preference weights providing scores that could match both rank 1 and 2 are w that is larger than 0.2.



(A) (B)

Figure 4-2. Trajectory scores varied by preference weights for two sets of trajectories (A) a full set $\{\tau_{[i]}\}_{i=1}^{10}$where the feasible preference weights $W_{[1]} = \{0.1 \le w \le 1\}$ and (B) a set after removing rank 1 $\{\tau_{[i]}\}_{i=2}^{10}$ where the feasible preference weights $W_{[2]} = \{0.2 \le w \le 1\}$

4.2.3. Optimization problem

Our problem has the objectives of 1) finding feasible preference weights $\boldsymbol{w}$ that provides scores matching with the individual trajectory rankings and 2) finding the optimal policies where each policy π_j^* is optimal based on each of the feasible preference weights. We formulate the problem as an optimization problem given one set of ranked trajectories $\{\tau_{[i]}\}_{i=1}^{M}$. All trajectories are generated from the same unknown policy π and each trajectory has a time horizon of T. The space of preference weights $\mathcal{W}$ is finite where $\mathcal{W} = \{\boldsymbol{w_j}\}_{j=1}^{N}$.

Optimization problem (P):

$$\max_{j\in[N]}\{\boldsymbol{w}^T\,\boldsymbol{\mu}(\pi_j^*)\}$$

subject to

$$\pi_j^* = \underset{\pi}{\operatorname{argmax}}\, E\left[\sum_{t=0}^{T}\gamma^t\, R(\boldsymbol{w_j}, s_t)|\pi\right], \qquad \forall \boldsymbol{w_j} \in \mathcal{W} \qquad (1)$$

$$sc_{[1]}\left(\boldsymbol{w}, \widehat{\boldsymbol{\mu}}_{[\boldsymbol{1}]}(\pi)\right) = \max_{i\in[M]}\left\{\boldsymbol{w}^T\widehat{\boldsymbol{\mu}}_{[i]}(\pi)\right\}, \qquad \forall \boldsymbol{w} \in \mathcal{W} \qquad (2)$$

$$sc_{[2]}\left(\boldsymbol{w}, \widehat{\boldsymbol{\mu}}_{[\boldsymbol{2}]}(\pi)\right) = \max_{i\in[2,\dots,M]}\left\{\boldsymbol{w}^T\widehat{\boldsymbol{\mu}}_{[i]}(\pi)\right\}, \qquad \forall \boldsymbol{w} \in \mathcal{W} \qquad (3)$$

$$\vdots$$

$$sc_{[M-2]}\left(\boldsymbol{w}, \widehat{\boldsymbol{\mu}}_{[\boldsymbol{M-2}]}(\pi)\right) = \max_{i\in[M-2,M-1,M]}\left\{\boldsymbol{w}^T\widehat{\boldsymbol{\mu}}_{[i]}(\pi)\right\}, \qquad \forall \boldsymbol{w} \in \mathcal{W} \qquad (4)$$

$$sc_{[M-1]}\left(\boldsymbol{w}, \widehat{\boldsymbol{\mu}}_{[\boldsymbol{M-2}]}(\pi)\right) = \max_{i\in[M-1,M]}\left\{\boldsymbol{w}^T\widehat{\boldsymbol{\mu}}_{[i]}(\pi)\right\}, \qquad \forall \boldsymbol{w} \in \mathcal{W} \qquad (5)$$

Our objective function represents the highest policy values for each value of $\boldsymbol{w}$ as represented by the black line in Figure 4-1. The constraints (2) to (5) reduce the space of $\mathcal{W}$ and result in the preference weights that could match all trajectory ranks.

4.2.4. Algorithm

To learn the preference weight $\boldsymbol{w}$ accurately, we create two algorithms for the problem with the preference dimension K of 2. We keep track of only one component of the preference weights w^1, as w^2 is always $1 - w^1$. The two algorithms are 1) Exhaustive search and 2) Heuristic search. Both algorithms iteratively reduce the preference weight space by obtaining several sets of ranked trajectories from an individual and remove the preference weights that do not provide the correct rankings. We assume that we have access to the first set of trajectories generated by an expert using the unknown policy π_e, which is optimal for the reward function using the unknown expert's preference $\boldsymbol{w_e}$.

The exhaustive search considers all possible preference weights in the space and generates their corresponding optimal policies. Then, it keeps only the preference weights that could match the correct rankings by solving the optimization problem P. In the next step, we calculate the difference or "gap" v between the policy values of the feasible preference weights at the boundaries of feasible region (w_{min} and w_{max}). The algorithm terminates if the gap v is smaller than a threshold ε or the number of iterations has reached the maximum iterations. Otherwise, we obtain another set of ranked trajectories generated from the policy, which is optimal for the selected

feasible weights (w_{min} or w_{max}) and solve the optimization problem P. The detail of the exhaustive algorithm can be found in Algorithm 1.

For the heuristic search, we create two storage sets for keeping the preference weights that can and cannot match the rankings, namely, Match and Mismatch. The algorithm iteratively updates the two sets and the feasible region boundary w_{min}^1 and w_{max}^1. In each iteration, the algorithm draws a preference weight w^1 by sampling without replacement from the preference weight space. After we obtain w^1, w^1 is checked whether it can match all trajectory rankings and is assigned to either "Match" or "Mismatch". We define that w^1 matches the rankings if it can provide trajectory scores that can be sorted into a ranked trajectory set with the same order as the individual's rankings. We use rules for reducing the feasible region boundary based on "Match" and "Mismatch" (Algorithm 2 and Figure 4-3), and rules for sampling a new w^1 while the algorithm is not terminated and when w_{min}^1 and w_{max}^1 have not changed after two iterations (Algorithm 2 and Figure 4-4). The algorithm terminates if the gap v is smaller than a threshold ε or the number of iterations reach the maximum iterations. Otherwise, we generate a new set of trajectories using the latest w^1 that matches the rankings. The detail of the heuristic search can be found in Algorithm 2.

Algorithm 1: Exhaustive search

Define: $\boldsymbol{w_p}$ = true individual preference where $\boldsymbol{w_p} = \left[w_p^1, 1 - w_p^1\right]^T$, $w_{min}^1 = 0.0$, $w_{max}^1 = 1.0$, $iter = 1$, $\mathcal{W} = [w_{min}^1 : k : w_{max}^1]$ where k = the increment between w_j and w_{j+1}, $maxIter$ = the maximum number of iterations

1. Obtain a set of M trajectories from expert $\{\tau_i\}_{i=1}^M$and ranks from an individual $\left\{\tau_{[i]}\right\}_{i=1}^M$, which is ranked by $\boldsymbol{w_p}$.
2. Solve the optimization problem P for all w and obtain a set of feasible solution $\mathcal{W}'$ and their corresponding policies. Update $\mathcal{W} \leftarrow \mathcal{W}'$.
3. Update $w_{min}^1 \leftarrow \min_w \mathcal{W}$ and obtain π_{min}
   Update $w_{max}^1 \leftarrow \max_w \mathcal{W}$ and obtain π_{max}
4. Let $\boldsymbol{w_{min}} = [w_{min}^1, 1 - w_{min}^1]^T$and $\boldsymbol{w_{max}} = [w_{max}^1, 1 - w_{max}^1]^T$. Calculate the policy value gap v as follows:
   $$v = |\boldsymbol{w_{min}}^T \widehat{\boldsymbol{\mu}}(\pi_{min}) - \boldsymbol{w_{max}}^T \widehat{\boldsymbol{\mu}}(\pi_{max})|$$
   where $\widehat{\boldsymbol{\mu}}(\pi_{min})$ is estimated from 1000 simulated trajectories using policy π_{min}.
   $\widehat{\boldsymbol{\mu}}(\pi_{max})$ is estimated from 1000 simulated trajectories using policy π_{max}.
5. If $v \leq \varepsilon$ or $iter \geq maxIter$:
   Terminate the algorithm.
6. Select w from $\{w_{min}^1, w_{max}^1\}$ that provides the highest policy value and obtain its corresponding optimal π_w.
7. Generate a new set of M trajectories $\{\tau'_i\}_{j=1}^M$ using the corresponding π_w and get ranks from an individual $\left\{\tau'_{[i]}\right\}_{i=1}^M$, which is ranked by $\boldsymbol{w_p}$.
8. Let $iter \leftarrow iter + 1$. Go back to step 2 and update the constraints by estimating $\widehat{\boldsymbol{\mu}}_{[i]}(\pi)$ from $\tau'_{[i]}$.

Algorithm 2: Heuristic search

Define: $\boldsymbol{w_p}$ = true individual preference where $\boldsymbol{w_p} = \left[w_p^1, 1 - w_p^1\right]^T$, $w_{min}^1 = 0.0$, $w_{max}^1 = 1.0$, $iter = 1$, $\mathcal{W} = [w_{min}^1 : k : w_{max}^1]$ where k = the increment between w_j and w_{j+1}, $maxIter$ = the maximum number of iterations, Match = {}, Mismatch = {}

1. Obtain a set of M trajectories from expert $\{\tau_i\}_{i=1}^M$and ranks from an individual $\left\{\tau_{[i]}\right\}_{i=1}^M$, which is ranked by $\boldsymbol{w_p}$.
2. Initialize $w_{iter}^1 \leftarrow$ Uniformly sampling from $\mathcal{W}$, $w_{iter}^2 \leftarrow 1 - w_{iter}^1$
3. Let $\boldsymbol{w_{iter}} = [w_{iter}^1, w_{iter}^2]^T$ and Update $\mathcal{W}$ by removing w_{iter}^1
4. If Match $\neq \emptyset$:
   Check whether each weight in "Match" matches the ranking of $\left\{\tau_{[i]}\right\}_{i=1}^M$
   If not: move the mismatched weight to "Mismatch"
5. Calculate trajectory scores for each trajectory τ_i using $sc\left(\boldsymbol{w_{iter}}, \widehat{\boldsymbol{\mu}}_{(i)}(\pi)\right)$ and let $\left\{\tau_{[i]_{w_{iter}}}\right\}_{i=1}^M$denote ranked trajectories according to the scores.

6. If $\boldsymbol{w_{iter}}$ matches the ranks; $\left\{\tau_{[i]_{w_{iter}}}\right\}_{i=1}^{M} = \left\{\tau_{[i]}\right\}_{i=1}^{M}$:
   Add w_{iter}^1 to "Match"
   Else:
   Add w_{iter}^1 to "Mismatch"
   If Match $= \emptyset$:
   $w_{iter}^1 \leftarrow$ Uniformly sampling from $\mathcal{W}$, $w_{iter}^2 \leftarrow 1 - w_{iter}^1$
   Update $\mathcal{W}$ by removing w_{iter}^1
   Go back to Step 4
7. If the size of Match ≥ 1 and the size of Mismatch ≥ 1 (see Figure 4-3):
   If size of Mismatch = 1:
   If $w_{Mismatch} \geq \max_{m \in \text{Match}} \{w_m^1\}$: $w_{max} \leftarrow w_{Mismatch}$
   Else: $w_{min} \leftarrow w_{Mismatch}$
   Else:
   If $\min_{m \in \text{Match}} \{w_m^1\} > \max_{u \in \text{Mismatch}} \{w_u^1\}$: $w_{min} \leftarrow \max_{u \in \text{Mismatch}} \{w_u^1\}$
   Else if $\min_{u \in \text{Mismatch}} \{w_u^1\} > \max_{m \in \text{Match}} \{w_m^1\}$: $w_{max} \leftarrow \max_{m \in \text{Match}} \{w_m^1\}$
   Else:
   $w_{min} \leftarrow \max\left\{w \in \text{Mismatch} \cap w \leq \min_{m \in \text{Match}} \{w_m^1\}\right\}$
   $w_{max} \leftarrow \min\left\{w \in \text{Mismatch} \cap w \geq \min_{m \in \text{Match}} \{w_m^1\}\right\}$
8. Clear Mismatch: Mismatch = {}
9. Solve for the optimal policy π_{w_min} and π_{w_max} for $\boldsymbol{w_{min}}$ and $\boldsymbol{w_{max}}$, respectively
10. Calculate the policy value gap v as follows:
$$v = \left|\boldsymbol{w_{min}}^T \hat{\mu}\left(\pi_{w\_min}\right) - \boldsymbol{w_{max}}^T \hat{\mu}\left(\pi_{w\_max}\right)\right|$$
    where $\hat{\boldsymbol{\mu}}(\pi_{min})$ is estimated from 1000 simulated trajectories using policy π_{min}.
    $\hat{\boldsymbol{\mu}}(\pi_{max})$ is estimated from 1000 simulated trajectories using policy π_{max}.
11. If $v \leq \varepsilon$ or $i \geq maxIter$:
    Terminate the algorithm
12. Solve for the optimal policy π_w for the latest $\boldsymbol{w}$ that matches the rankings
13. Generate a new set of trajectories $\{\tau'_i\}_{i=1}^{M}$ using π_w and get ranks from an individual.
14. Update $\mathcal{W} \leftarrow w_{max} \geq \mathcal{W} \geq w_{min}$
15. Sampling w_{iter}^1 from $\mathcal{W}$ according to the following rules (see Figure 4-4):
    If $\mathcal{W} = \emptyset$: Let midpoint = median of Match,
    If $w_{iter-1}^1 \geq$ midpoint:
    $w_{iter}^1 \leftarrow$ Sampling from U[min{Match}, midpoint], see Figure 4-4 (B)
    Else: $w_{iter}^1 \leftarrow$ Sampling from U[midpoint, max{Match}], see Figure 4-4 (A)

    Else: Let midpoint = median of $\mathcal{W}$
    If there is no change in only w_{max} between $iter$ and $iter - 1$:
    $w_{iter}^1 \leftarrow$ Sampling from U[midpoint, w_{max}], see Figure 4-4 (A)

    Else if there is no change in only w_{min} between $iter$ and $iter - 1$:
    $w_{iter}^1 \leftarrow$ Sampling from U[w_{min}, midpoint], see Figure 4-4 (B)

Else If there is no change in w_{min} and w_{max} between $iter$ and $iter-1$:
Let $q \leftarrow$ randomly sampling between 0 and 1
If $q \leq 0.5$: $w_{iter}^1 \leftarrow$ Sampling from U[midpoint, w_{max}], see Figure 4-4 (A)
Else: $w_{iter}^1 \leftarrow$ Sampling from U[w_{min}, midpoint], see Figure 4-4 (B)

Else: $w_{iter}^1 \leftarrow$ Sampling from U[w_{min}, w_{max}] , see Figure 4-4 (C)

Let $w_{iter}^2 = 1 - w_{iter}^1$

16. Update space $\mathcal{W}$ by removing w_{iter}^1
17. Let $iter \leftarrow iter + 1$ and go back to Step 4

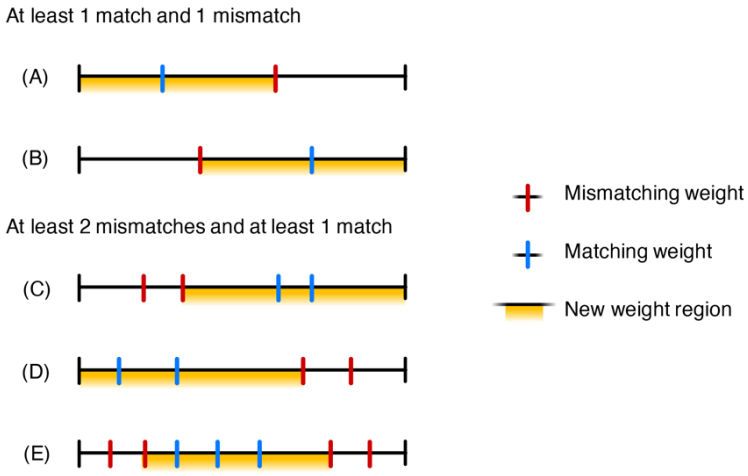


Figure 4-3. Rules for reducing the preference weight space based on matching and mismatching weights (Step 6 of Algorithm 2). The yellow region represents the new reduced space. The rules are for all possible scenarios: (A) and (D) mismatched weight is larger than the matching weights, (B) and (C) mismatched weight is smaller than the matching weights, and (E) mismatched weights are between matching weights.

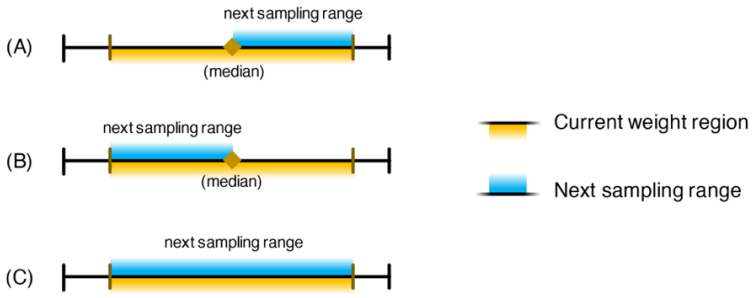


Figure 4-4. Rules for sampling preference weights when w_{min} or w_{max} has not changed after two consecutive iterations (Step 14 of Algorithm 2). Rules are applied as follows: 1) use (A) when only w_{max} has not changed, 2) use (B) when only w_{min} has not changed, 3) use either (A) or (B) when both have not changed, 4) use (C) otherwise.

4.3. Results

In this section, we setup a simulation environment for evaluating our proposed algorithms and present the results for a baseline, followed by a one-way sensitivity analysis to determine how various parameters affect the performance of each algorithm compared to the established baseline.

4.3.1. Simulation experimental setup

In order to evaluate the performance of the proposed algorithms, we conducted numerical experiments using an example problem in a Markov decision process. States in the problem are defined as (h, m) where h represents a health state and m represents the treatment state. The space of h is $\{1, 2, 3, D\}$ where D is an absorbing state (i.e., death) and the space of m is $\{1, 2, 3\}$. The action space A is all available treatments $\{1, 2, 3\}$. At each time period t, an individual can transition from h_t to h_{t+1} according to the selected action a_t where $a_t = m_{t+1}$ with a probability of $p(h_{t+1}|h_t, a_t)$. The transition probabilities are shown in Table 4-1. We consider a reward function with two functions, which are health utility and side effect burden from treatment according to Table 4-2.
The reward function is expressed as follows:

$$R(\boldsymbol{w}, s_t) = \begin{cases} w^1 \times u(h_t) + (1 - w^1) \times \{1 - d(m_t)\}, & \text{if } h_t \neq D \\ 0, & \text{if } h_t = D \end{cases}$$

Table 4-1. The transition probability matrix by each action

Action a_t	1	2	3
Transition Matrix	$\begin{bmatrix} 0.3 & 0.6 & 0.1 & \delta \\ 0.1 & 0.2 & 0.7 & \delta \\ 0.1 & 0.2 & 0.7 & \delta \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0.4 & 0.1 & \delta \\ 0.2 & 0.4 & 0.4 & \delta \\ 0.2 & 0.4 & 0.4 & \delta \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.7 & 0.2 & 0.1 & \delta \\ 0.3 & 0.5 & 0.2 & \delta \\ 0.3 & 0.5 & 0.2 & \delta \\ 0 & 0 & 0 & 1 \end{bmatrix}$

where $\delta \approx 0.00003$. Each row of the matrix represents the current health state h_t where the first row is $h_t = 1$ and the last row is $h_t = D$. Each column of the matrix represents the health state in the next period h_{t+1} where the first column is $h_{t+1} = 1$ and the last column is $h_{t+1} = D$. For example, if $h_t = 1$ and $a_t = 2$, then the probability of transitioning from $h_t = 1$ to $h_{t+1} = 2$ is 0.4, the element in the first row and the second column for $a_t = 2$.

Table 4-2. Health utility and disutility from treatment by each state

h_t	1	2	3	4
Health utility $u(h_t)$	0.9	0.6	0.3	0
m_t	1	2	3	
Disutility $d(m_t)$	0.17	0.2	0.25	

With our setup, we first establish a baseline by evaluating each algorithm's performance in estimating the preference weight of 50 different cases, whereas each case is consisted of a true individual weight w_p, an initial expert weight w_e (Table 4-3), and other baseline parameters (Table

4-4). These cases are selected such that there are combinations where w_e is close to w_p and combinations where w_e is very different from w_p. Using the obtained baseline, we perform one-way sensitivity analysis on the parameters including the size of preference weight space, the number of ranked trajectories, the length of time horizon, and the policy gap threshold. We present results from 50 cases using the following performance metrics: 1) the absolute error between w_{min} and w_p and between w_{max} and w_p, 2) the absolute optimality gap (Eq. 6), 3) the number of iterations, and 4) the algorithm runtime in the following subsections.

To calculate the absolute optimality gap, we used the following equations:

$$\text{absolute gap} = \begin{cases} \left| \boldsymbol{w_p}^T \widehat{\boldsymbol{\mu}} \left(\pi_{w_p} \right) - \boldsymbol{w_p}^T \widehat{\boldsymbol{\mu}} \left(\pi_{w_{min}} \right) \right| & \text{where } \pi_{w_{min}} \text{ is optimal policy for } R(\boldsymbol{w_{min}}, s) \\ \left| \boldsymbol{w_p}^T \widehat{\boldsymbol{\mu}} \left(\pi_{w_p} \right) - \boldsymbol{w_p}^T \widehat{\boldsymbol{\mu}} \left(\pi_{w_{max}} \right) \right| & \text{where } \pi_{w_{max}} \text{ is optimal policy for } R(\boldsymbol{w_{max}}, s) \end{cases} \quad (6)$$

where π_{w_p} is an optimal policy for $R(\boldsymbol{w_p}, s)$.

Our optimality gap is a theoretical metric, as we applied a true preference weight $\boldsymbol{w_p}$ to both terms, while in practice, the true preference weight $\boldsymbol{w_p}$ is unknown to a decision maker. As we assumed that an individual only receives reward based on their true preference weight $\boldsymbol{w_p}$, the first term of Eq. 6 represents a benchmark of the policy value, specifically the ground truth value of the optimal policy. The second term is the policy value that an individual receives when the policy is optimal based on the reward parametrized by $\boldsymbol{w_{min}}$ or $\boldsymbol{w_{max}}$.

Table 4-3. Fifty cases formed by combinations of true individual weight w_p and initial expert weight w_e

Case no.	w_p	w_e	Case no.	w_p	w_e	Case no.	w_p	w_e
1	0.1	0.1	21	0.1	0.5	41	0.1	0.9
2	0.2		22	0.2		42	0.2	
3	0.3		23	0.3		43	0.3	
4	0.4		24	0.4		44	0.4	
5	0.5		25	0.5		45	0.5	
6	0.6		26	0.6		46	0.6	
7	0.7		27	0.7		47	0.7	
8	0.8		28	0.8		48	0.8	
9	0.9		29	0.9		49	0.9	
10	1		30	1		50	1	
11	0.1	0.3	31	0.1	0.7			
12	0.2		32	0.2				
13	0.3		33	0.3				
14	0.4		34	0.4				
15	0.5		35	0.5				
16	0.6		36	0.6				
17	0.7		37	0.7				
18	0.8		38	0.8				
19	0.9		39	0.9				
20	1		40	1				

Table 4-4. Baseline parameters

Parameter	Baseline	Varied values
Increment interval in preference weight space (k)	0.05	0.01 and 0.1
Number of ranked trajectories per set (M)	10	3 and 5
Time horizon (T)	10	30 and 50
Max iterations	20	-
Policy gap threshold (ε)	0.5	0.25 and 1

4.3.2. Numerical results for baseline

At baseline, both algorithms perform well regarding the absolute errors of preference weights and the absolute optimality gap. Both algorithms successfully converge w_{min} and w_{max} towards the true w_p with small errors for all 50 cases. According to Figure 4-5, the exhaustive search provides the mean absolute errors (MAE) of 0.10 and 0.05 for w_{min} and w_{max}, respectively. The heuristic search provides MAE of values either 0.05 or 0.10 for both w_{min} and w_{max}. For the optimality gap, both algorithms provide values less than 0.05 for both optimal policies obtained from w_{min} and w_{max} (Figure 4-6).

According to Figure 4-7, the exhaustive search terminates within 4 iterations on average, while the mean number of iterations for the heuristic search is either 1 or 2. Ideally, we desire the search algorithm to terminate as early as possible because each additional iteration increases the number of times we have to inquire the rankings from an individual.

On average, the exhaustive search has an algorithm runtime of 8.80 seconds per case, while the heuristic search has a runtime of 2.37 seconds per case. This shows that the heuristic search is at least 3 times faster than the exhaustive search. We can observe that the main reason for the slower runtime in the exhaustive search is the exhaustive exploration of feature expectations and optimal policies for all possible preference weights in the space.

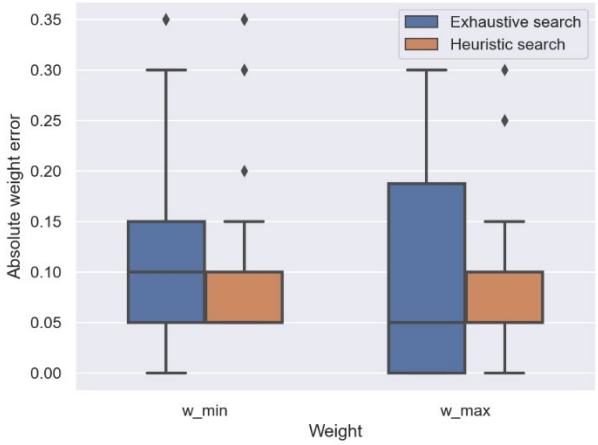


Figure 4-5. Boxplot of the mean absolute errors between w_{min} and w_p, and between w_{max} and w_p.

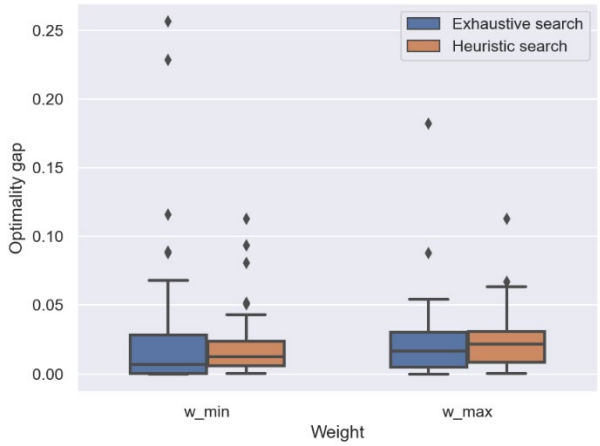


Figure 4-6. Boxplot of the mean absolute errors of the optimality gap between $\pi^*_{w_{min}}$ and $\pi^*_{w_p}$, and between $\pi^*_{w_{max}}$ and $\pi^*_{w_p}$

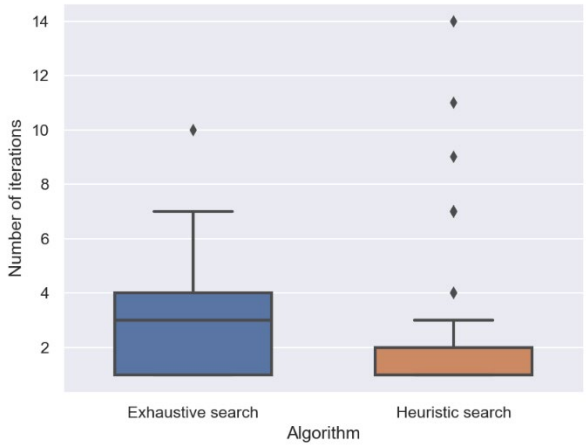


Figure 4-7. The number of iterations to reach termination for each algorithm.

4.3.3. Numerical results for one-way sensitivity analysis

To understand how our algorithms behave regarding changes in parameters, we performed one-way sensitivity analysis on 1) the size of preference weight space specified by the increment interval k, 2) the number of ranked trajectories per set M, 3) the time horizon T, and 4) the policy gap threshold ε. As we vary one parameter, we keep all other parameters at baseline as shown in Table 4-4. All results are reported as the mean over 50 cases as shown in Table 4-5 and Table 4-6 for the exhaustive search and the heuristic search algorithms, respectively. The following presents the effects on metrics resulting from the varied parameters in the sensitivity analysis:

Effects on MAE of preference weights and optimality gap

According to the results, we observe that changes in the size of preference weight space (k) have no significant effect on the MAE of the resulting preference weights and optimality gap in the exhaustive search algorithm. The exhaustive search algorithm will always explore all possible

preference weights in the space, therefore, the convergence of w_{min} and w_{max} to w_p should be stable regardless of an increase in the size of space. With a smaller size of the preference weight space, the MAE of preference weights in the heuristic search seems to increase. However, when $k = 0.01, 0.05, 0.1$, the smallest possible MAEs are 0.01, 0.05, and 0.1, respectively. Therefore, we conclude that there is no significant change in the MAE of preference weights. In addition, varying the size of preference weight space has no effect on the optimality gap.

As the number of ranked trajectories decreases, we observe an increase in MAE of preference weights for the exhaustive search algorithm, while the heuristic search algorithm provides similar MAEs. We found no significant change in optimality gap in both algorithms except for $M = 3$ in the exhaustive search.

For both algorithms, we found no changes in the MAE of preference weights as the time horizon is varied. As the gap threshold increases, both algorithms perform worse in the accuracy of providing the correct preference weights as the algorithms terminate too early. The optimality gap is not affected by changes in time horizon and the gap threshold for both algorithms.

Effects on algorithm runtime and number of iterations

The exhaustive search algorithm performs similarly in the number of iterations as the increment interval (k) increases, whereas the heuristic search performs worse. We observe that the iterations required increases by approximately 5 and 3 times when the number of ranked trajectories is 3 compared to the number of ranked trajectories at baseline ($M = 10$) for the exhaustive and the heuristic search, respectively. For both algorithms, a smaller number of ranked trajectories per set leads to more iterations required to terminate. We believe that when the number of ranked trajectories per set is small, there is a higher chance for one preference weight value to match all ranks correctly compared to when the number of ranked trajectories per set is large. Consequently, more iterations are required to remove mismatched preference weights from the space. As the time horizon per trajectory is longer, we observe that more iterations are required to terminate for both algorithms except for when $T = 50$ in the heuristic search. When the gap threshold decreases, both algorithms require more iterations to terminate in order to fulfill the stricter terminating condition.

Lastly, our results show that runtimes of the exhaustive search algorithm are affected by only the size of the weight space and the time horizon. By decreasing the increment interval, the size of preference weight space increases leading to longer runtimes for the exhaustive search, which is expected. As the space size increases by 10 times ($k = 0.1$ to $k = 0.01$), the mean runtime increases by at least 10 times as well (4.4 seconds to 44.3 seconds). Similarly, an increase in the time horizon per trajectory leads to a longer runtime. As the time horizon increases by 5 times ($T = 10$ to $T = 50$), the mean runtime increases by at least 3 times (8.8 seconds to 31.0 seconds). As for the heuristic search algorithm, we found that the runtime is roughly proportional to the number of iterations, therefore, factors affecting the runtime include increases in increment interval, increases in time horizon duration, decreases in the number of ranked trajectories, and decreases of the gap threshold. Overall, the heuristic search has a shorter runtime compared to the exhaustive search.

Table 4-5. The average numerical results over 50 cases for each parameter using exhaustive search.

Parameter	Value	MAE of w_{min}	MAE of w_{max}	Absolute gap of $\pi_{w_{min}}$	Absolute gap of $\pi_{w_{max}}$	Mean iterations	Mean runtime (seconds)
Increment interval in preference weight space (k)	0.01	0.11	0.08	0.03	0.02	2.88	44.27
	0.05	0.11	0.08	0.02	0.02	3.75	8.80
	0.1	0.11	0.08	0.03	0.03	2.84	4.60
Number of ranked trajectories per set (M)	3	0.22	0.18	0.17	0.03	15.06	9.06
	5	0.13	0.10	0.03	0.03	9.26	9.30
	10	0.11	0.08	0.02	0.02	3.75	8.80
Time horizon (T)	10	0.11	0.08	0.02	0.02	3.75	8.80
	30	0.09	0.09	0.04	0.03	7.38	20.79
	50	0.09	0.07	0.07	0.04	9.54	31.01
Policy gap threshold (ε)	0.25	0.07	0.05	0.01	0.02	5.74	8.73
	0.5	0.11	0.08	0.02	0.02	3.75	8.80
	1	0.18	0.17	0.11	0.03	1.38	8.83

Table 4-6. The average numerical results over 50 cases for each parameter using heuristic search.

Parameter	Value	MAE of w_{min}	MAE of w_{max}	Absolute gap of $\pi_{w_{min}}$	Absolute gap of $\pi_{w_{max}}$	Mean iterations	Mean runtime (seconds)
Increment interval in preference weight space (k)	0.01	0.03	0.03	0.02	0.02	1.10	1.01
	0.05	0.09	0.08	0.02	0.02	2.59	2.37
	0.1	0.14	0.11	0.04	0.02	6.34	5.63
Number of ranked trajectories per set (M)	3	0.10	0.11	0.05	0.03	5.84	5.47
	5	0.11	0.10	0.03	0.02	5.58	5.13
	10	0.09	0.08	0.02	0.02	2.59	2.37
Time horizon (T)	10	0.09	0.08	0.02	0.02	2.59	2.37
	30	0.09	0.08	0.07	0.03	8.78	17.75
	50	0.09	0.06	0.07	0.04	6.94	21.54
Policy gap threshold (ε)	0.25	0.07	0.07	0.02	0.02	8.32	7.41
	0.5	0.09	0.08	0.02	0.02	2.59	2.37
	1	0.13	0.08	0.06	0.02	1.32	1.21

4.4. Discussion

We considered a problem of learning an individual's preference and finding an optimal policy under a Markov decision process where the reward function is consisted of two objectives weighted by the individual's unknown preference. Two search algorithms, exhaustive search and heuristic search, were developed for our problem. We found that both algorithms successfully provide the preference weight ranges that converge to the true weight with small errors. Both algorithms perform well over all 50 simulation cases at the baseline in optimality gaps and requires a low number of rankings queries. Comparing the two algorithms, the heuristic search can efficiently and accurately learn an individual's preference under a short runtime.

The two algorithms have their own advantages and disadvantages, and therefore, selecting which one to use depends on the objectives and a problem's characteristics. If the preference weight space of the problem in question is small, then the exhaustive search algorithm is more suitable as the heuristic search algorithm needs more iterations to terminate in this specific situation. If the goal is to obtain accurate preference weights, learn the preference weights within the smallest number of iterations or runtime, or using the smallest number of ranked trajectories, then using the heuristic search is preferable.

The heuristic search may seem to be more suitable for most cases, aside from when the threshold gap is very small as heuristic search requires more iterations to terminate compared to exhaustive search. However, upon further inspection, we found that the heuristic search can terminate earlier than reported, as the preference weight ranges have already converged to values close to true weight. The high number of iterations is, in fact, due to the gap threshold being too small compared to the difference between the policy values at the current preference ranges. Consequently, additional rules for terminating the algorithm might be needed. For example, one rule could be to count the number of consecutive iterations with the preference weight ranges remaining constant and compare this number to a prespecified threshold. If the number exceeds the threshold, then the algorithm terminates. As such, addressing this issue will require further numerical experiments.

To apply our algorithms in practice, we need to first estimate an individual's transition probability matrix for each available action. If there exist an individual's sequential observations of health states and actions, we can use the maximum likelihood estimation, a common approach for transition matrix estimation [96]. If the individual's observations are not available, we can estimate a population transition matrix using the same technique from an existing observational study of a population on our interested action or treatment. Otherwise, we need to conduct an observational study. One key step of our algorithm is to request an individual to provide a complete ranking of a set of trajectories according to their preference. The trajectories may be difficult to interpret depending on the specific problem. Therefore, showing a trajectory in an easily interpretable way is necessary. For example, instead of showing a trajectory numerically, we can show a curve of possible states over time, such as a person's bodyweight, and the frequency of actions, such as caloric intake.

Our work can be generalized broadly to decision-making problems under a Markov decision process with the following assumptions: 1) a reward function is a linear combination of two objectives, and is parametrized by unknown preference weights, 2) an individual's transition

probability is known, 3) complete rankings of multiple sets of trajectories can be obtained from the individual. In the area of personalized medicine, our algorithms can be applied to problems that preference affects patient outcomes. For example, patients with depression who achieve better health outcomes after being on their preferred treatment [49], and patients with schizophrenia who have to decide among competing objectives such as benefits gained and side effect burdens [108,109]. In just-in-time adaptive intervention, we believe that our framework can be applied to behavioral management problems similar to MyBehavior [121], a smartphone application tracking an individual's physical activity and diet behavior and provides suggestions that incorporates an individual's preference from daily questionnaire. For instance, using mobile sensing systems, we can first estimate the individual's transition probability from their sequences of states such as calories, choices of exercise activities, and choices of diet using a maximum likelihood estimation [96]. Next, we provide a questionnaire by showing multiple curves of potential changes in calories with the corresponding actions to take and asking them to rank all the curves. Then, we learn the preference using the heuristic search using data from the rankings and provide optimal sequential actions based on the learned preference and their transition probability.

This study has several limitations. A complete ranking could be burdensome for some individuals, especially when the number of trajectories is large. To address this issue, other elicitation methods [107] can be applied such as 1) asking for the most and least preferred trajectories if the number of trajectories required is small (Best-worst choice) and 2) sequentially asking to choose the most preferred trajectory among two trajectories from the same set (Binary choice-sequence). In addition, our heuristic search algorithm is developed for problems with a reward function that is a linear combination of two objectives. To extend beyond two objectives, we could represent the preference weight space as a convex hull with the summation of all weight elements equals 1 and keep track of the new convex hull as mismatched weights are removed from the original convex hull.

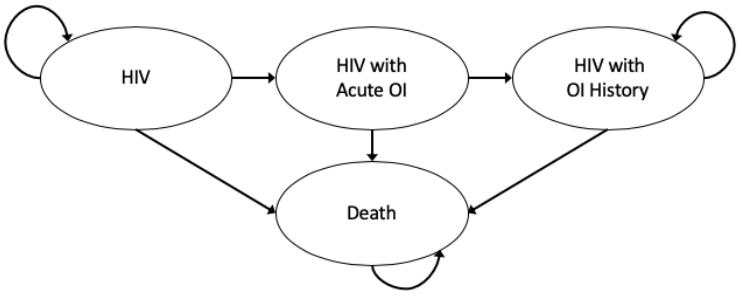
In conclusion, we developed the exhaustive search and the heuristic search for the problem of learning an individual's preference weight and the corresponding optimal policy under Markov decision process given an expert's trajectories and complete rankings of trajectories from an individual. The algorithms were evaluated on their performance using numerical experiments. Both of our proposed algorithms can successfully search for the preference weight ranges that converge to the true weight, with the heuristic search algorithm performing well in all performance metrics and is more flexible compared to the exhaustive search algorithm. Our work introduces a new algorithm for estimating an individual's preference given a complete ranking and can be applied to various healthcare and consumer choice problems.

Appendix A

I. HIV model

We developed a microsimulation model that simulates the natural disease progression and treatment status in HIV-infected children in the presence of drug resistance. The model follows the disease progression of a cohort of chronic HIV-infected children from the age of 0 to 13 using a monthly cycle. Each individual is tracked on CD4%/CD4 cell count, viral load, antiretroviral therapy (ART) regimens, treatment duration, treatment failure status, opportunistic infections (OI) history, and drug resistance status. As the main immunologic measure changes from CD4% to CD4 cell count at age 5, we incorporated a translation of CD4% to CD4 cell count to our model, which were explained in Appendix A-II. The model was used in simulating 500,000 sub-Saharan HIV-infected children with ART initiation from the age of 3 years and provided the outputs of health outcomes and costs for several treatment strategies.

1. Disease progression



Appendix A Figure 1. HIV disease progression

HIV state transition

The children population in our model have an initial state of chronic HIV or "HIV" state. At each period, they can remain in the "HIV" state or transition to the "HIV with Acute OI" state if a clinical event occurs (Appendix A Figure 1.). After 30 days of a clinical event occurrence, children transition from the "HIV with Acute OI" to the "HIV with OI History" state. All states can transition to the death state. The state with "OI" is associated with an increase in the probability of death stratified by CD4 level.

CD4 dynamics

At the beginning of the simulation, we assigned CD4% to the pre-ART HIV-infected children of aged 0 by randomly sampling from the initial CD4% distribution, which is assumed to be Normal distribution with the mean and standard deviation of 44.2% and 10.0%, respectively (Table 2-1). Without effective treatment, the children suffered a decline in CD4 level each month with a constant rate stratified by age for children aged < 5 and a decline rate stratified by viral load level was applied to children aged ≥ 5 (Table 2-2). The equations for calculating CD4%/CD4 cell count for each time period are presented in Appendix A Table 1. Note that CD4% is as the main immunologic measure for children aged ≤ 5, while CD4 cell count is used in children aged ≥ 5.

Appendix A Table 1. The CD4 progression for the children without effective ART

Age	CD4 progression
Children aged < 5	CD4%(t) = CD4%$(t-1)$ × (1- a rate of CD4% decline)
Children aged ≥ 5	CD4 cell count(t) = CD4 cell count$(t-1)$ - CD4 cell count decline

where CD4%(t) is CD4% at period t and CD4 cell count(t) is CD4 cell count level at period t.

For those who are on effective ART (no virologic failure), their CD4 level is increased at a constant rate. The increase rate is stratified by ART regimen and duration on ART with a maximum durations of 48 months (Appendix A Table 2). The CD4 progression on effective ART is calculated as follows:

$$\text{CD4 } (t) = \text{CD4 } (t-1) + \text{an increase in CD4}$$

where CD4(t) is CD4% or CD4 cell count at period t.

Appendix A Table 2. Parameters relevant to recovery from being on effective ART

Parameters	**Values**	**Source**
Monthly CD4% for children age < 5 year		
NNRTI first 6 months, after 6 months	2.3%, 0.7%	Calibrated
PI first 6 months, after 6 months	2.0%, 0.4%	Calibrated
DTG first 6 months, after 6 months	2.3%, 0.7%	Assumed
Monthly CD4 cell count increase for children age ≥ 5 year		
NNRTI/PI/DTG first 6 months	67.3 cells/μL	[35]
NNRTI/PI/DTG after 6 months	3.4 cells/μL	[35]
Number of effective months for CD4	48	
Reduction in event risk for patients on ART		
RRR of clinical event	0.88	Calibrated
RRR of mortality risk for age	0.97	Calibrated

NNRTI: non-nucleoside reverse transcriptase inhibitor; DTG: dolutegravir; PI: protease inhibitor; RRR: relative risk reduction.

Viral load dynamics

In our model, initial viral load is assigned to children at age 5. Children who have not been on ART were assigned a viral load sampling from a truncated Normal distribution with a mean viral load of 5.1 $\log_{10}$ copies per mL and a standard deviation of 0.84 $\log_{10}$ copies per mL, where the lower and upper bounds are 2.5 and 7.5 $\log_{10}$ copies per mL, respectively [122,123]. On the other hand, children on ART were assigned with the lower bound of viral load, 2.5 $\log_{10}$ copies per mL.

When children are on effective ART, viral load declines each month with a maximum number of six months. A decrease in viral load is 0.27 logs per month [124,125]. In this model, we assumed the lowest viral load level has a value of 1.6 $\log_{10}$ copies per mL or 40 copies/mL [126]. If an

individual is lost to follow-up or virologic failure occurs, the viral load level bounces back to the highest record level of that individual.

Opportunistic infections

In children aged < 5, the risk of developing OI (or clinical events) depends on age where the age group of 6–59 months has a higher risk of developing OI, compared to the age group of < 6 months (Table 2-1). In children aged ≥ 5, the risk of developing OI also depends on age where the age group of 5–9 years has a higher risk of developing OI, compared to the age group of 10–14 years (Table 2-2). During 30 days of clinical events (acute OI), there is a drop in CD4 cell count regardless of ART status for children aged ≥ 5. For children who are on effective ART, there is a reduction in risks of developing OI (Appendix A Table 2). The equation for calculating the probability of OI for children on effect ART are as follows:

$$\text{Probability of OI on ART}(age) = (1 - \text{RRR OI}) \times \text{Probability of OI no ART}(age)$$

The probability of OI in children aged > 5 are calculated using relative risk and the probability of OI in children aged ≤ 5. We first obtained the relative risk (RR) for children aged 5–9 years and aged 10–14 years by comparing the average rate of WHO3 and WHO4 [127] at each of the older age group to that of the age group of 2–4 years, respectively (Appendix A Table 3). Then, we multiplied each RR value (0.56 and 0.25) to the average monthly rate of OI in children aged < 5 (4.05%) and converted the rate back to the risk as shown in the last column of Appendix A Table 4.

Appendix A Table 3. Risk of clinical event for children aged 5–9 years and 10–14 years relative to children aged 2–4 years

Age (years)	WHO 3 IR	WHO 4 IR	Average IR	Relative risk
	per 100 PY			
2–4	9.40	3.20	6.30	reference
5–9	4.60	2.50	3.55	0.56
10–14	1.80	1.30	1.55	0.25

Appendix A Table 4. Derivation of the probability of OI for children aged 5–9 years and 10–14 years

Children aged < 5	Monthly risk of OI	Monthly rate of OI		Children aged < 5	Monthly risk of OI	Monthly risk of OI
0–6 months	3.32%	0.034		5–9	0.023	2.28%
> 6 months	4.62%	0.047		10–14	0.010	1.00%
	Average	0.041				

The monthly risk of OI for children aged < 5 is our model parameters from the calibration.

Mortality

The risks of mortality are stratified by a history of OI, age, and CD4% level (Table 2-1 and Table 2-2). Lower CD4%/CD4 Cell count is associated with a higher risk of mortality. The risk of morality with no history of OI and with history of OI is higher in the younger age groups. Within

30 days of clinical events (acute OI), there is a relatively high risk of mortality regardless of CD4% level, compared to the risk of morality with no history of OI. For those who are on effective ART, there is a reduction in risks of mortality (Appendix A Table 2). The equation for calculating the probability of death for children on effect ART are as follows:

$$\text{Probability on ART}(age, CD4) = (1 - \text{RRR mortality}) \times \text{Probability no ART}(age, CD4)$$

The probability of death in children aged > 5 are calculated using similar approach to the probability of OI in children aged > 5. The relative risk is shown in Appendix A Table 5 with the mortality incidence rates obtained from [127]. We obtained an average monthly rate of death by CD4% from the probability of death for children aged 25–36 months, 37–48 months, and 49–60 months (Table 2-1). Then, we multiply the average rate (Appendix A Table 6) with relative risk for each age group (0.32 and 0.18). Next, we converted the multiplied rate to the monthly risk as shown in the last two rows of Appendix A Table 6.

Appendix A Table 5. Mortality risk for children aged 5–9 years and 10–14 years relative to children aged 2–4 years

Age (years)	Mortality IR (per 100 PY)	Relative risk
2–4	3.40	reference
5–9	1.10	0.32
10–14	0.60	0.18

Appendix A Table 6. Derivation of the probability of death for children aged 5–9 years and 10–14 years

Average monthly rate of death	with no OI history			with OI history			within 30 days of OI
	<15%	15-24%	≥25%	<15%	15-24%	≥25%	
25–60 months	0.0073	0.0082	0.0080	0.0556	0.0182	0.0087	0.1176
Average monthly rate of death× RR	with no OI history			with OI history			within 30 days of OI
	<200	200-499	≥500	<200	200-499	≥500	
5–9 years	0.0023	0.0026	0.0026	0.0180	0.0059	0.0028	0.0380
10–14years	0.0013	0.0014	0.0014	0.0098	0.0032	0.0015	0.0208
Probability of death	with no OI history			with OI history			within 30 days of OI
	<200	200-499	≥500	<200	200-499	≥500	
5–9 years	0.23%	0.26%	0.26%	1.78%	0.59%	0.28%	3.73%
10–14years	0.13%	0.14%	0.14%	0.98%	0.32%	0.15%	2.05%

2. Treatment progression

Once children reached age of 3, ART was initiated if HIV was diagnosed regardless of CD4 level. After ART initiation, the model assumed that children were on effective ART until viral load is tested. During a period that children received viral load testing, children were tested whether they failed on the treatment or not, where a probability of virologic failure depends on ART regiments and drug resistance status (Table 2-3). The maximum number of months that virologic failure can occur is 12 months after every ART initiation. If children do not experience virologic failure after 12 months on ART, the ART regimen is assumed to be effective for the rest of the time horizon. When a virologic failure occurred, children have a chance to switch to the second-line ART, which is specified by the probability of switching to second-line ART (Table 2-3).

Pretreatment drug resistance

A proportion of children population in our model is initialized to have pretreatment drug resistance (PDR) to non-nucleoside reverse-transcriptase inhibitors (NNRTI) at birth. For those who have PDR to NNRTI, there is an increase in a risk of virologic failure on NNRTI-based ART, compared to those without PDR (Table 2-3). Note that the PDR status can be known only through PDR testing.

Treatment strategies

We simulated five ART strategies: 1) standard of care with empiric efavirenz (NNRTI)-based first-line ART (*status quo*); 2) PDR testing to guide choice of initial ART regimen (*PDR testing*); 3) increasing rate of switching to protease inhibitor (PI)-based second-line ART (*improved switching*); 4) introducing dolutegravir (DTG)-based ART as the empiric first-line regimen (*DTG status quo*); and 5) DTG-based ART as the empiric first-line regimen combined with increasing rate of switching to second-line ART (*DTG improved switching*). In all strategies, we assumed that only two ART regimens are available, namely, NNRTI-based or DTG-based first-line ART and PI-based second-line ART.

In *status quo* and *improved switching* strategies, children initiate ART with the empiric NNRTI-based first-line ART regardless of their PDR status. If virologic failure occurs, children on failed treatment has a chance to switch to PI-based ART. For those who have PDR to NNRTI, a risk of virologic failure on NNRTI-based ART is higher than those without PDR. In *PDR testing* strategy, children were tested on their PDR status at the age of 3 prior to ART initiation. If PDR is detected, children initiate ART with PI-based ART, if not detected, they initiate ART with NNRTI-based ART. In *DTG status quo* and *DTG improved switching* strategies, children initiate ART with DTG-based first-line ART regardless of their PDR status and switch to PI-based ART if virologic failure occurs. Parameters relevant to each ART strategy are presented in Table 2-3.

For all strategies, we assumed that children initiate a treatment if they meet one of the two conditions: 1) Have recent acute OI and 2) Receive positive results. For those who are on a second-line ART, we assumed that they continue on the treatment regardless of virologic failure due to a benefit from continuing on a failed treatment comparing to discontinuing a treatment [25].

3. Assumptions for CD4, viral load, and HIV testing

CD4 and viral load testing
Children received CD4 testing at ART initiation and when virologic failure is detected to assess for risk of OI. For those who are on ART, they received viral load testing 6 months after ART initiation and every 12 months thereafter.

HIV testing
In our model, children are tracked on the HIV testing status where the status can be positive, negative, or never tested. All children begin with the status of never tested. Their status can be changed to positive by one of the two criteria: 1) Eligible to test and 2) Identified by OI. The first criterion refers to children who have not been tested before or who have never tested positive or are up for retest (retest occurs after 12 months). The second criterion refers to children who have recent acute OI.

II. HIV parameters

1. Model calibration

Calibration is used to ensure that our microsimulation model for HIV-infected patients has well-estimated input parameters, which resulted from comparing the model outputs to empirical data [128]. We first obtained observational targets from existing studies and selected several parameters to vary in the calibration process. Each parameter was varied under either Uniform or Normal distribution. We then created 20,000 different sets of parameters and each parameter set is with 200,000 patients. Mean squared error is used in measuring the goodness-of-fit, and the 50 best sets of parameters are selected. Detailed explanation of our calibration process is described in the following subsections.

Observational Targets
We calibrated our model against four observational targets: 1) the observational survival outcomes from the UNAIDS, 2) P1060 observed mortality rate and observed OI rate [129], 3) the statistics of CD4% and CD4 cell count of HIV-infected patients at age 5 (Appendix A Table 4). To obtain the three outputs from simulation, several conditions are applied to the model according to each target.

The UNAIDS survival outcomes is for the population of HIV-infected children with no treatment from age 0 to 60 months. The study population included more than 1,300 children from eight sub-Saharan African countries. The survival outcomes are reported for children at age of 6 months, 12 months, 24 months, 36 months, 48 months, and 60 months. We obtained the values for age of 6 months, 12 months, and 24 months from [23] and the remaining values were obtained from [22].

We obtained the P1060[7] observed mortality rate from Ciaranello et al. [22], which is 3.29/100PY over a median follow-up of 72 weeks. The P1060 observed OI rate was calculated from the average over three clinical events from the same study: WHO3 (9.30/100PY), WHO4 (0.73/100PY), and

[7] P1060 study is a clinical trial for comparing antiviral responses to NNRTI-based and PI-based therapy in HIV-infected infants who have and have not received single dose nevirapine.

TB (5.60/100PY). The average value of 5.21/100PY was used as our calibration target. The population of IMPAACT[8] P1060 trail is HIV-infected children presenting to care at 12 months of age. The children population is randomized to two arms, nevirapine (NVP) and lopinavir (LPV) arms. In NVP arm, children initiate ART with NNRTI-based ART as the first-line ART and PI-based ART as the second-line ART. In LPV arm, children initiate ART with PI-based ART as the first-line ART and NNRTI-based ART as the second-line ART. Each arm has a probability of viral suppression over 12 months as shown in Appendix A Table 7. Note that our model-generated mortality rate and OI rate are calculated from the simulation over the first two years after ART initiation.

In our simulation, CD4% is converted to CD4 cell count when children reached the age of 5. Therefore, we investigated the values of CD4% and CD4 cell count in children at age 5 from the existing literature that focus on the population of HIV-infected children in sub-Saharan Africa with the mean age close to 5 who received no treatment (Appendix A Table 8). Thus, the median, the 25^{th} percentile, and the 75^{th} percentile of CD4% and CD4 cell count of children at age of 5 (a total of six values) were used as the calibration targets. For these targets, we considered a cohort of HIV-infected children with no treatment and followed them from age of 0 to 5.

CD4 Translation

The immunological measure of HIV progression in children aged < 5 and children aged ≥ 5 are CD4% and CD4 cell count, respectively. CD4% is the ratio of CD4 count to lymphocytes count. However, there is a lack of data on the absolute value of lymphocytes and CD4 in HIV-infected children in Kenya, consequently, we could not calculate CD4 cell count directly.

To translate CD4% to CD4 cell count of children in our model, we assumed that CD4 cell count follows a Normal distribution as CD4% is assumed to follow a normal distribution. We approximated CD4 cell count of each individual in our simulated population by matching the i^{th} percentile of CD4% to the i^{th} percentile of CD4 cell count when the population reaches age of 5. For example, a 5-year-old child from the population has CD4% of 12, which is 50^{th} percentile in the CD4% distribution, was assigned CD4 cell count of 370 cells/μl, which is 50^{th} percentile in the CD4 cell count distribution.

Parameters included in the calibration

Three categories of parameters (19 parameters in total) were selected to vary in our calibration process, which are CD4 related, mortality related, and ART related parameters. Parameters in three categories, their sampling distribution, and the ranges of variation are specified in Appendix A Table 9. In our model, we assumed there are three types of mortality risk: risk of mortality with no history of OI, risk of mortality within 30 days of OI, and risk of mortality after 30 days of OI. As the risk of mortality with no history of OI and after 30 days of OI are stratified by CD4 level and age groups, we consider additional 13 calibration inputs, mortality multipliers, which are stratified by age groups. By multiplying the mortality multipliers to the risk of mortality by CD4 level (after converting to the monthly rate), we obtain the risk of mortality by both CD4 level and age. Combining all related parameters, there are 32 parameters in total in our calibration.

[8] IMPAACT: International Maternal Pediatric Adolescent AIDS Clinical Trials Group

Appendix A Table 7. The probability of viral suppression over 12 months from [22].

	NVP arm	LPV arm
First-line ART	72% (NNRTI)	86% (PI)
Second-line ART	70% (PI)	71% (NNRTI)

Appendix A Table 8. CD4% and CD4 cell count of HIV-infected children at age close to 5

Mean of Age (years)	Median of CD4%		Median of CD4 cell count (cells/μl)		Source
7.1 (3.6, 4.2)	11.9	(5.8, 17.5)	356	(132, 603)	[130]
6.4 (3.5, 9.6)	12	(7, 17)	-		[131]
4.83 (1.7, 9.08)	14	(8.9, 20)	381	(180, 734)	[132]
7.6 (4.5, 11)	17	(7.5, 28)	397	(183, 800)	[133]
6.3 (2.4, 9.7)	12	(6, 17)	-		[134]
Calibration targets	12	(6, 17)	370	(160, 700)	

The provided values are in the format of mean (min, max) and median (25^{th} percentile, 75^{th} percentile).

Goodness-of-fit measures

Since our calibration has multiple targets, we use mean squared error (MSE) to measure the goodness-of-fit. For each target, we calculated the mean squared error and scaled by the target ranges (Appendix A Table 11) to eliminate the differences in the target unit. For example, there are 5 data points from the UNAIDS survival outcomes, which are the proportion of children alive at different ages. The ranges of each data point were $\pm$10% of its corresponding target. Other target boundary are summarized in Appendix A Table 11

Calibration results

With MSE as the goodness-of-fit, we ranked all 20,000 sets of 32 parameters according to the mean MSE over all targets and selected the 50 best-fitted sets. The minimum and maximum of the mean MSE of 50 sets achieved from the calibration process are 0.006 and 0.03. In this appendix, we reported the model outcomes of the four sets of parameters with smallest mean MSE in Appendix A Figure 2 and Appendix A Table 12. The model-generated survival outcomes by the four sets of parameters have small deviations from the UNAIDS survival outcomes (Appendix A Figure 2).

Appendix A Table 9. Parameters related to disease and treatment progression included in calibration.

Input parameters for calibration	Sampling Distribution	Normal distribution parameters (Mean, SD)	Ranges for varying	Source
CD4				
Mean of CD4 cell count at age 5	Uniform		301-559[a]	
SD of CD4 cell count at age 5	Uniform		280-520[a]	
Initial mean of CD4% at age 0	Uniform		42-50%	[22]
Initial SD of CD4% at age 0	Uniform		9.40-14.10%	[22]
CD4% Monthly decline by age				
≤ 3 months	Uniform		3.00-8.00%	[22]
4 months - 5 years	Uniform		0.30-0.70%	[22]
Mortality				[22]
Monthly probability of death with no history of clinical event by CD4% level				
CD4% < 15	Normal	0.40%, 0.02%	0.32-0.49%	
CD4% 15-24	Normal	0.40%, 0.02%	0.33-0.49%	
CD4% >= 25	Normal	0.40%, 0.02%	0.32-0.49%	
Probability of death within 30 days of clinical events	Normal	3.10%, 0.05%	2.88-3.31%	
Monthly probability of death with history of clinical event (> 30 days post-event) by CD4% level				
CD4% < 15	Normal	2.40%, 0.05%	2.22-2.61%	
CD4% 15-24	Normal	0.80%, 0.03%	0.70-0.92%	
CD4% >= 25	Normal	0.40%, 0.02%	0.30-0.50%	
ART				[35]
Relative risk reduction in event risk for patients on ART				
Opportunistic infection	Uniform		0.85-1	
Mortality	Uniform		0.90-1	
CD4% monthly increase in children aged < 5				
NNRTI in the first 6 months	Normal	2.20%, 0.17%	2.20-2.86%	
NNRTI after 6 months	Normal	0.70%, 0.05%	0.70-0.91%	
PI in the first 6 months	Normal	1.90%, 0.09%	1.90-2.28%	
PI after 6 months	Normal	0.40%, 0.02%	0.40-0.48%	

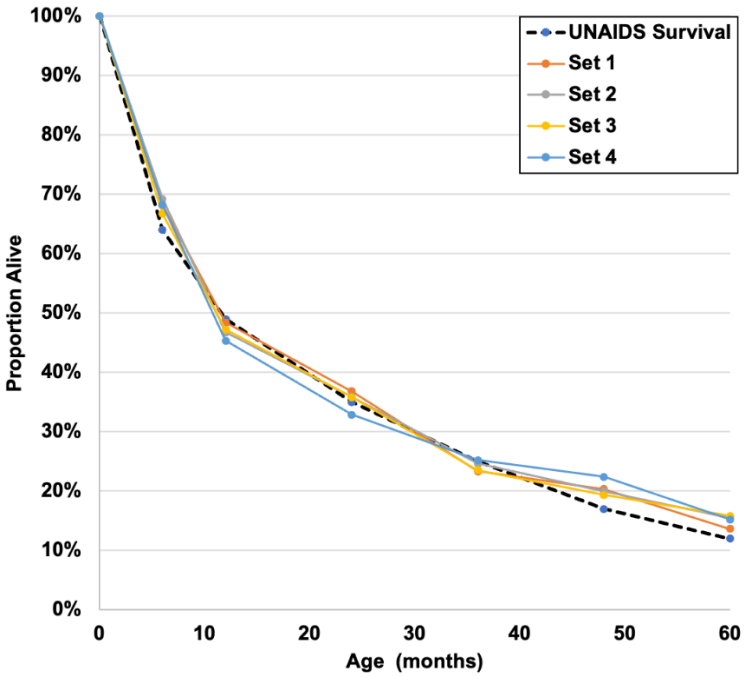
a) The ranges of CD4 cell count mean and SD is calculated by solving a system of equations with two unknowns (Mean and SD) using the 25th percentile and the 75th percentile in Appendix A Table 4. The two equations are 1) Mean – 0.675SD = 160 and 2) Mean + 0.675SD = 700. The solutions to the equations are: Mean = 430 and SD = 400. Then, the ranges for CD4 cell count mean are 430±30% and CD4 cell count SD are 400±30%.

Appendix A Table 10. Mortality multipliers included in calibration

Multipliers for calibration	Sampling Distribution	Ranges for varying
Multipliers for monthly probability of death with no history of clinical event		
Age 0-6 months	Uniform	1-20
Age 7-12 months	Uniform	1-20
Age 13-24 months	Uniform	0.5-5
Age 25-36 months	Uniform	0.5-5
Age 37-48 months	Uniform	0.2-2
Age 49-60 months	Uniform	0.2-2
Multipliers for probability of death within 30 days of clinical events	Uniform	0.5-5
Multipliers for monthly probability of death with history of clinical event (> 30 days post-event)		
Age 0-6 months	Uniform	1-20
Age 7-12 months	Uniform	1-20
Age 13-24 months	Uniform	0.5-5
Age 25-36 months	Uniform	0.5-5
Age 37-48 months	Uniform	0.2-2
Age 49-60 months	Uniform	0.2-2

Appendix A Table 11. Target boundary.

Calibration target	Minimum	Maximum
UNAIDS survival outcomes	-10% of the target	+10% of the target
CD4%	1%	40%
CD4 cell count (cells/μl)	100	1200
Mortality rate (rate per 100 PY)	2.0	5.0
OI rate (rate per 100 PY)	0.73	9.30



Appendix A Figure 2. Observed UNAIDS survival outcomes and model-generated survival outcomes by four parameter sets

Appendix A Table 12. Model outputs from the four parameter sets

Model outputs	**Targets**	**Set 1 (0.006)**	**Set 2 (0.007)**	**Set 3 (0.008)**	**Set 4 (0.009)**
CD4% at age 5					
Median	12	9.23	12.22	13.71	10.07
25th percentile	6	2.33	5.25	7.16	2.72
75th percentile	17	15.34	18.59	20.00	16.66
CD4 cell count at age 5					
Median	370	496.13	492.39	492.21	495.11
25th percentile	160	281.71	280.23	279.17	283.25
75th percentile	700	729.62	722.85	720.93	726.29
P1060 observed rate					
Mortality rate per 100PY	3.29	3.34	3.14	3.64	3.18
OI rate per 100PY	5.21	5.15	5.02	4.66	5.10

Each set is corresponding to its rank based on MSE in parenthesis i.e. set 2 has the second least value of MSE

2. Pretreatment drug resistance (PDR) prevalence

According to Boerma et al., the meta-analysis on PDR of children in sub-Saharan Africa analyzed and reported the prevalence by the prevention of mother-to-child transmission (PMTCT) exposure and by ages [5]. The PDR is defined as the resistance to non-nucleoside reverse transcriptase inhibitor (NNRTI) or nucleoside reverse transcriptase inhibitors (NRTI) agents, which are agents contained in NNRTI-based ART. According to their analysis, the pooled PDR prevalence of children aged ≤ 12 years was 42.7% among those exposed to antiretroviral drugs for the PMTCT and 12.7% among those with no exposure. The PDR prevalence of children aged < 3 years was 40% and children aged ≥ 3 years was 17.6% regardless of the PMTCT exposure status. In our model assumptions, we do not include the children status of PMTCT exposure. In addition, the population of interest is sub-Saharan African HIV-infected children, who initiate ART at the age of 3 years. Therefore, we assume that PDR prevalence is approximately 18% in our base-case analysis.

According to the 2019 WHO Drug Resistance Report, the estimated PDR prevalence among newly diagnosed infants in sub-Saharan Africa to NNRTI agents alone is ranged from >10% to >30% for those with no or unknown PMTCT exposure, and is ranged from >30% to >70% for those with PMTCT exposure. In addition, the PDR prevalence to NRTI agents alone is ranged from <5% to > 20%. To account for the varied PDR prevalence across countries in sub-Saharan Africa, we consider the ranges of are 5% and 30% in our one-way sensitivity analysis.

3. Probability of virological failure

<u>NNRTI-based ART with and without PDR as a first-line ART</u>

According to the meta-analysis by Boerma et al., the rate of viral suppression in HIV-infected children from low- and middle-income countries (LMIC), including Africa and Asia, receiving first-line ART after 12 months was 72.7% [28]. As some of the studies included in analyzing the reported rate of viral suppression did not specify children's status of PMTCT exposure, we assumed that the reported rate was the overall rate of viral suppression for a population of children with and without PMTCT exposure. To estimate the probabilities of virologic failure of NNRTI-based ART in children by PDR status, we used the overall rate of viral suppression, the base-case PDR prevalence, and the odds ratio of virologic failure in children on NNRTI-based ART with PDR compared to without PDR.

We solved a system of equations with two unknowns (equations (9) to (11)); $\mathrm{VF}_{\mathrm{NNRTI\text{-}PDR}}$ and $\mathrm{VF}_{\mathrm{NNRTI\text{-}NOPDR}}$. Using the input values of 72.7%, 18%, and 7.5 for Rate of $\mathrm{suppression}_{\mathrm{NNRTI}}$, PDR, and $\mathrm{OR}_{\mathrm{NNRTI}}$, respectively, we obtained the base-case values of $\mathrm{VF}_{\mathrm{NNRTI\text{-}PDR}}$ and $\mathrm{VF}_{\mathrm{NNRTI\text{-}NOPDR}}$ as 64.1% and 19.2%, respectively.

$$\mathrm{VF}_{\mathrm{NNRTI}} = 1 - \text{Rate of } \mathrm{suppression}_{\mathrm{NNRTI}} \quad (9)$$

$$\mathrm{VF}_{\mathrm{NNRTI}} = \mathrm{PDR} \times \mathrm{VF}_{\mathrm{NNRTI\text{-}PDR}} + (1 - \mathrm{PDR}) \times \mathrm{VF}_{\mathrm{NNRTI\text{-}NOPDR}} \quad (10)$$

$$\mathrm{OR}_{\mathrm{NNRTI}} = \{\mathrm{VF}_{\mathrm{NNRTI\text{-}PDR}}/(1 - \mathrm{VF}_{\mathrm{NNRTI\text{-}PDR}})\}/\{\mathrm{VF}_{\mathrm{NNRTI\text{-}NOPDR}}/(1 - \mathrm{VF}_{\mathrm{NNRTI\text{-}NOPDR}})\} \quad (11)$$

where VF_{NNRTI} is the virologic failure rate of NNRTI-based ART in children, $VF_{NNRTI\text{-}PDR}$ is the virologic failure rate of NNRTI-based ART in children with NNRTI-associated PDR, and $VF_{NNRTI\text{-}NOPDR}$ is the virologic failure rate of NNRTI-based ART in children without PDR.

To obtain the ranges for sensitivity analyses for probabilities of virologic failure of NNRTI-based ART in children with PDR, we solve the above equations using the same inputs except for the odds ratio, which now have the values of of 2 and 15. According to Kityo et al. [7], the odds ratio of 15 was the point estimate for all mutations (NRTI or NNRTI) and either on efavirenz or nevirapine. As we did not include nevirapine-based ART as one of the NNRTI-based ART regimens, we considered 15 as the upper bound. The lower bound of 2 was half of the study's lower bound of a 95% confidence interval.

PI-based ART as a first-line ART

There is a limited study on children aged ≥ 3 years initiating ART with PI-based ART as PI-based ART was not one of the recommended initial treatments for HIV-infected children aged ≥ 3 years according to the guidelines [135]. Due to the lack of available information, we assumed that the probability of virologic failure of PI-based as a first-line ART is the same as the probability of virologic failure of NNRTI-based as a first-line ART for children without NNRTI-associated PDR, which is consistent with [17,136].

DTG-based ART as a first-line ART

Due to the lack of available information of DTG-based ART in children, we estimated the probability of virologic failure of DTG-based ART in children using adult information. According to Dugdale et al. [137], the estimated probabilities of viral suppression in South African women without PDR to NNRTI after 48 weeks on NNRTI-based ART and after 48 weeks on DTG-based ART are 91% and 96%, respectively. Based on the two probabilities, the ratio of the odds for virologic failure of NNRTI-based ART compared to DTG-based ART in women without PDR to NNRTI is 2.37.

To estimate the probability of virologic failure of DTG-based ART in children, we assumed that the odds ratio for virologic failure of NNRTI-based ART compared to DTG-based ART in children is the same as in adults ($OR_{NNRTI\ to\ DTG} = 2.37$). Solving the equation (12) with $VF_{NNRTI\text{-}NOPDR}$ of 19.2%, we obtained the probability of virologic failure of DTG-based ART in children with a value of 9.10%, which is our base-case value.

$$OR_{NNRTI\ to\ DTG} = \{VF_{NNRTI\text{-}NOPDR}/(1 - VF_{NNRTI\text{-}NOPDR})\}/\{VF_{DTG}/(1 - VF_{DTG})\} \quad (12)$$

PI-based ART as a second-line ART

According to a multicenter analysis of children in Asia and sub-Saharan Africa, they found that 16.4% of children receiving PI-based second-line ART over 24 months have virologic failure with a 95% confidence interval of 13.9% and 19.4% [26]. Consequently, the probability of PI-based second-line ART after NNRTI-based first-line ART in our base-case analysis is 16.4%, and the 95% confidence intervals are used for our sensitivity analysis. Due to the lack of data availability, we assumed the probability of PI-based second-line ART after DTG-based first-line ART is the same as the probability of PI-based second-line ART after NNRTI-based first-line ART. To address the uncertainty of the value and account for a higher failure rate due to poor adherence,

we consider the larger range of 13.9% to 40.0% for our sensitivity analysis when DTG-based ART is the first-line ART.

4. Probability of switching to second-line ART by 1 year after failure

We assumed the probability of switching to second-line ART is 40%. This assumption is based on pediatric studies on a switch to a second-line ART after virologic failure on a first-line ART. The study on the IeDEA Pediatric Cohort with more than 90% of children are from African countries found that 4763 children had first-line ART failure after a median follow-up of 19.5 months with 45.4% found by virologic criterion [20]. After a median follow-up of 14.3 months, only 20.8% were switched to second-line ART. According to the study on South African children, they reported that only 38% out of 252 children who have a first-line treatment failure were switched to second-line after had at least 1 year of follow-up [138].

5. Cost of ART regimens

We retrieved the costs and dosages of antiretroviral treatment from the Global Fund antiretroviral drugs pricing reference [4] and the Pediatric guidelines from HHS[9] [135]. We considered two potential first-line ART regiment for children from age 3 to 14. The NNRIT-based ART are ABC + 3TC + EFV[10] and AZC + 3TC + EFV[11]. For the PI-based ART, we considered AZT + 3TC + LPV/r[12] or ABC + 3TC + LPV/r.

Using data shown in Appendix A Table 13 and Appendix A Table 14, we calculated the cost of NNRTI-based ART per person per year by averaging costs from the weight category, which has the value of $133 and $113 for ABC + 3TC + EFV and AZC + 3TC + EFV, respectively. Therefore, the NNRTI-based ART cost per person per year on average is $123.

For PI-based ART, we used the information shown in Appendix A Table 15 and Appendix A Table 16. Using the weighted average, we have the cost per person per year of $300 and $280 for ABC + 3TC + LPV/r and AZC + 3TC + LPV/r, respectively. Therefore, the cost of PI-based ART per person per year on average is $290.

Appendix A Table 13. Data for calculating the cost of ABC + 3TC + EFV, NNRTI-based ART

<table>
<tr><th rowspan="2">Weights (kg)</th><th colspan="2">ABC + 3TC</th><th colspan="2">EFV</th></tr>
<tr><th>Available Product</th><th>Tabs per day</th><th>Available Product</th><th>Tabs per day</th></tr>
<tr><td>10 – 13.9</td><td rowspan="3">Abacavir/Lamivudine 120/60mg tablet dispersible 30 ($3.49)</td><td>2</td><td rowspan="4">Efavirenz 200mg tablet 30 ($1.50)</td><td>1</td></tr>
<tr><td>14 – 19.9</td><td>2.5</td><td>1.5</td></tr>
<tr><td>20 – 24.9</td><td>3</td><td>1.5</td></tr>
<tr><td>25 – 34.9</td><td>Abacavir/Lamivudine 600/300mg tablet 30 ($9.20)</td><td>1</td><td>2</td></tr>
</table>

[9] HHS: United States Department of Health and Human Services
[10] ABC: Abacavir; 3TC: Lamivudine; EFV: Efavirenz;
[11] AZT: Zidovudine;
[12] LPV/r: Lopinavir/Ritonavir

Appendix A Table 14. Data for calculating the cost of AZT + 3TC + EFV, NNRTI-based ART

Weights (kg)	AZT + 3TC		EFV	
	Available Product	Tabs per day	Available Product [4]	Tabs per day
10 – 13.9	Lamivudine/Zidovudine 150/300mg tablet 60 ($5.25)	1	Efavirenz 200mg tablet 30 ($1.50)	1
14 – 19.9		1		1.5
20 – 24.9		1.5		1.5
25 – 34.9		2		2

Appendix A Table 15. Data for calculating the cost of ABC + 3TC + LPV/r, PI-based ART

Weights (kg)	ABC + 3TC		LPV/r	
	Available Product	Tabs per day	Available Product	Packs per month
10 – 13.9	Abacavir/Lamivudine 120/60mg tablet dispersible 30 ($3.49)	2	Lopinavir/Ritonavir 80/20mg/ml oral solution 60ml*5 ($30.82)	2
14 – 19.9		2.5		2.5
20 – 24.9		3		3
25 – 34.9	Abacavir/Lamivudine 600/300mg tablet 30 ($9.20)	1		3

Appendix A Table 16. Data for calculating the cost of AZC + 3TC + LPV/r, PI-based ART

Weights (kg)	AZT + 3TC		LPV/r	
	Available Product	Tabs per day	Available Product	Packs per month
10 – 13.9	Lamivudine/Zidovudine 150/300mg tablet 60 ($5.25)	1	Lopinavir/Ritonavir 80/20mg/ml oral solution 60ml*5 ($30.82)	2
14 – 19.9		1		2.5
20 – 24.9		1.5		3
25 – 34.9		2		3

6. Cost of inpatient and outpatient cares

We obtained costs per inpatient bed-day and costs per outpatient visit for several countries in sub-Saharan Africa that represent a wide range of GDPs including South Africa [29], Ghana [31], Kenya [30], Zambia [33], and Uganda [32]. As the reported values are not in 2020 US$, we converted the costs to 2020 US$ using consumer price index (CPI) and use the average value as our base-case value. The derivations of the cost of inpatient care and the cost of outpatient are presented in Appendix A Table 17 and Appendix A Table 18. The equation for calculating the cost in to 2020 US$ is as follows:

$$\text{Cost in 2020 US\$} = \text{Cost in 'year reported' US\$} \times \text{CPI of 2020} / \text{CPI of 'year reported'}$$

Appendix A Table 17. Costs per inpatient bed-day for five countries in sub-Saharan Africa converted to average cost in 2020 US$

Countries	Year reported	Cost in year reported US$	CPI of year reported M01	CPI 2020 M01	Cost in 2020 US$	GDP in 2019
South Africa	2005	$82.95[a]	51.78	113.8	$182.30	$6,001
Ghana	2011	$39	42.26	113.9	$105.11	$2,202
Kenya	2011	$41	110.57	201.57	$74.74	$1,817
Zambia	2010	$20	105	246.72	$46.99	$1,305
Uganda	2011	$41	105.23	180.26	$70.23	$794
				Average	$95.88	

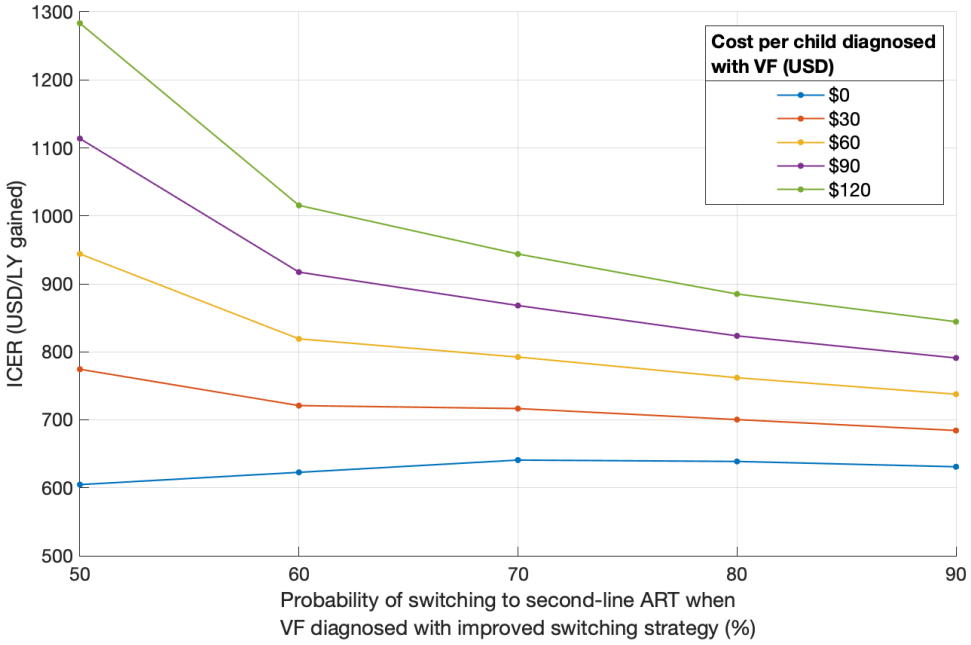
a) The value is calculated from the average over the costs of care from different hospital levels: primary = $60.89, secondary = $79.44, tertiary = $108.51.

Appendix A Table 18. Costs per outpatient visit for five countries in sub-Saharan Africa converted to average cost in 2020 US$

Countries	Year reported	Cost in year reported US$	CPI of year reported M01	CPI 2020 M01	Cost in 2020 US$	GDP in 2019
South Africa	2005	$31.46[a]	51.78	113.8	$69.14	$6,001
Ghana	2011	$14	42.26	113.9	$37.73	$2,202
Kenya	2011	$10	110.57	201.57	$18.23	$1,817
Zambia	2010	$8	105	246.72	$18.80	$1,305
Uganda	2011	$8	105.23	180.26	$13.70	$794
				Average	$31.52	

a) The value is calculated from the average over the costs of care from different hospital levels: primary = $20.90, secondary = $29.64, tertiary = $43.84.

III. Additional results



Appendix A Figure 3. Scenario analyses with a cost incurred from improving the rate of switching to second-line ART when the probability of virologic failure on PI-based second-line ART increases to 40% over 24 months.

Appendix B

I. Methodology

1. Proof for Proposition 1. The probabilities of transitioning to better states in P^m is higher than those of P.

Let u denote the health state at period t and let w denote the following heath states at period $t+1$ for all $t \in 1,2,..,T$:

$$w \in \begin{cases} \{1\}, & \text{if } u = 1 \\ \{1,2\}, & \text{Otherwise} \end{cases}$$

Let v denote the heath states at period $t+1$ where $v \neq w$.
We will show that $p^m(w|u) \geq p(w|u)$.

Consider the sum of probabilities of transitioning to better states $p^m(w|u)$

$$\begin{aligned} \sum_w p^m(w|u) &= 1 - p(D|u) - \sum_v p^m(v|u) = 1 - p(D|u) - \rho^m \sum_v p(v|u) \\ &= 1 - p(D|u) - \rho^m \left\{ 1 - p(D|u) - \sum_w p(w|u) \right\} \\ &= \{1 - \rho^m\} - \{1 - \rho^m\} p(D|u) + \rho^m \sum_w p(w|u) \\ &= \{1 - \rho^m\}\{1 - p(D|u)\} + \rho^m \sum_w p(w|u) \\ &= \sum_w r_w \{1 - \rho^m\}\{1 - p(D|u)\} + \rho^m p(w|u) \end{aligned}$$

where $r_w = \frac{p(w|u)}{\sum_w p(w|u)}$

For fixed w, we define $p^m(w|u)$ as follows:

$$p^m(w|u) = \rho^m p(w|u) + \frac{p(w|u)}{\sum_w p(w|u)} \times \{1 - \rho^m\}\{1 - p(D|u)\} \quad (1)$$

To show that $p^m(w|u) \geq p(w|u)$, we express $p(w|u)$ as follows:

$$p(w|u) = \rho^m p(w|u) + (1 - \rho^m) \times p(w|u) \quad (2)$$

The difference between equation (1) and (2) is in the second term of the right-hand-side of the equations and we can see that

$$\frac{1}{\sum_w p(w|u)} \times \{1 - \rho^m\}\{1 - p(D|u)\} \times p(w|u) \geq (1 - \rho^m) \times p(w|u)$$

$$\frac{1 - p(D|u)}{\sum_w p(w|u)} \geq 1$$

$$\frac{\sum_w p(w|u) + \sum_{v \neq D} p(v|u)}{\sum_w p(w|u)} \geq 1$$

$$1 + \frac{\sum_{v \neq D} p(v|u)}{\sum_{w} p(w|u)} \geq 1$$

Thus, this completes the proof that $p^m(w|u) \geq p(w|u)$.

The probabilities of transitioning to worse states in P^m are lower than those of P.
Let u denote the health state at period t and let v denote the following heath states at period $t+1$ for all $t \in 1,2,..,T$:

$$v \in \begin{cases} \{2,3\}, & \text{if } u = 1 \\ \{3\}, & \text{Otherwise} \end{cases}$$

The probability of transitioning to a worse state $p^m(v|u)$ is defined as follows:

$$\begin{aligned} p^m(v|u) &= \rho^m \times p(v|u), \qquad \rho^m \in [0,1] \\ &\leq p(v|u) \end{aligned}$$

2. Solution to the maximization problem of the log-likelihood function

Solving the maximization problem using Lagrange multipliers:
Let $f(\rho)$ denote the log-likelihood function.
Let λ_1, λ_2 denote the Lagrange multipliers for constraints (3).

$$L(\rho, \lambda_1, \lambda_2) = f(\rho) + \lambda_1(-\rho + \bar{\rho} - k\sigma) + \ \lambda_2(\rho - \bar{\rho} - k\sigma)$$

where $\lambda_1, \lambda_2 \geq 0$

Case I: Find the gradient of L over ρ and set the gradient to zero,

$$\begin{aligned} \nabla_\rho L(\rho, \lambda_1, \lambda_2) = &\frac{1}{\rho} \sum_{x_1 \neq D} \sum_{x_2 > x_1} N(x_1, x_2) + \frac{N(3,3)}{\rho} \\ &+ \sum_{x_1 \neq D} \sum_{x_2 \leq x_1, x_2 \neq 3} N(x_1, x_2) \times \left[\frac{p(x_1, x_2) - \rho q_{s_1 s_2}\{1 - p(x_1, D)\}}{[\rho p(x_1, x_2) + q_{x_1 x_2}\{1-\rho\}\{1 - p(x_1, D)\}]} \right] - \lambda_1 + \lambda_2 \\ &= 0 \end{aligned}$$

$$\begin{aligned} &\frac{1}{\rho}\left[\sum_{x_1 \neq D} \sum_{x_2 > x_1} N(x_1, x_2) + N(3,3) \right] + \sum_{x_1 \neq D} \sum_{x_2 \leq x_1, x_2 \neq 3} N(x_1, x_2) \times \left[\frac{p(x_1, x_2) - \rho q_{s_1 s_2}\{1 - p(x_1, D)\}}{[\rho p(x_1, x_2) + q_{x_1 x_2}\{1-\rho\}\{1 - p(x_1, D)\}]} \right] \\ &+ (\lambda_2 - \lambda_1) = 0 \end{aligned}$$

Therefore, the roots of the left-hand side of the above equation are among the optimal solutions.

Case II: Find the gradient of L over λ_1 and set the gradient to zero,

$$\nabla_{\lambda_1} L(\rho, \lambda_1, \lambda_2) = -\rho + \bar{\rho} - k\sigma = 0 \rightarrow \rho^* = \bar{\rho} - k\sigma$$

Therefore, $\rho^* = \bar{\rho} - k\sigma$ is one of the optimal solutions.

Case III: Find the gradient of L over λ_2 and set the gradient to zero,

$$\nabla_{\lambda_2} L(\rho, \lambda_1, \lambda_2) = \rho - \bar{\rho} - k\sigma = 0 \rightarrow \rho^* = \bar{\rho} + k\sigma$$

Therefore, $\rho^* = \bar{\rho} + k\sigma$ is one of the optimal solutions.

3. Proof of convexity of the log-likelihood function

$$\log p(O) = \sum_{x_1 \neq D} \left\{ N(x_1, D)\log p(x_1, D) + \sum_{x_2 > x_1} N(x_1, x_2)\log \rho p(x_1, x_2) + \sum_{x_2 \leq x_1, x_2 \neq 3} N(x_1, x_2)\log \left[\rho p(x_1, x_2) + q_{x_1 x_2}\{1-\rho\}\{1 - p(x_1, D)\}\right] \right\} + N(3,3)\log \rho p(3,3)$$

Let $g(\rho) = -\log \rho p(x_1, x_2)$ and $h(\rho) = -\log \left[\rho p(x_1, x_2) + r_{x_1 x_2}\{1 - p(x_1, D)\} \times (1-\rho)\right]$
$g(\rho)$ and $h(\rho)$ are twice differentiable, then we can find their Hessian $\nabla^2 g(\rho)$ and $\nabla^2 h(\rho)$

$$\nabla g(\rho) = -\frac{1}{\rho}, \qquad \nabla^2 g(\rho) = -\left(-\frac{1}{\rho^2}\right) = \frac{1}{\rho^2} \geq 0$$

Therefore, $g(\rho)$ is convex in ρ for $\rho \neq 0$.

$$\nabla h(\rho) = \frac{-\left[p(x_1, x_2) - q_{x_1 x_2}\{1 - p(x_1, D)\}\right]}{\rho p(x_1, x_2) + r_{x_1 x_2}\{1 - p(x_1, D)\} \times (1-\rho)}$$

$$\nabla^2 h(\rho) = \frac{\left[p(x_1, x_2) - q_{s_1 s_2}\{1 - p(x_1, D)\}\right]^2}{\left[\rho p(x_1, x_2) + r_{x_1 x_2} \times \{1 - p(x_1, D)\}(1-\rho)\right]^2} \geq 0$$

Therefore, $h(\rho)$ is convex in ρ.

$g(\rho)$ and $h(\rho)$ are convex, then $-g(\rho)$ and $-h(\rho)$ are concave. Summation of a concave function is still concave. Therefore, $\log p(H)$ is a concave function.

II. Additional results

1. Final model of Q-function approximated by FQI algorithm

a. FQI-Ridge

Ridge regression is a variation of linear regression. The difference between linear regression and ridge regression is the method of estimating the model coefficients. In linear regression, we use ordinary least squares or minimize the residual sum of squares between the observed targets in the dataset. In Ridge regression, we minimize a penalized residual sum of squares.

Linear regression:

$$\vec{y} = X\vec{w}$$

Coefficient estimation for linear regression:

$$\min_{w} \|X\vec{w} - \vec{y}\|_2^2$$

Coefficient estimation for ridge regression:

$$\min_{w} \|X\vec{w} - \vec{y}\|_2^2 + \alpha\|\vec{w}\|_2^2$$

where X is a feature matrix, $\vec{w}$ is a vector of coefficients, $\vec{y}$ is the observed target in the dataset, and α is regularization strength.

The coefficients estimated by ridge regression can be interpreted as how important each feature is. If a coefficient of one feature is close to zero, then that feature is less important in estimating our interested target. On the other hand, if the coefficient of one feature has a relatively large value, then that feature is considered more important in the target estimation.

To better understand our final model, we provide the top 10 most important features of our model instead of coefficients of each feature. The top 10 features of our base-case results for setup 1 and setup 2 can be found in Appendix B Table 1 and Appendix B Table 2, respectively. Note that our features have a dimension of 102 as one-hot encoding is used in converting categorical data to binary data. Therefore, we only report the top 10 features of our final model obtained from the last learning trial (the 10^{th} trial).

Appendix B Table 1. The top 10 most important features of the final FQI-Ridge model in setup 1

Rank	Fast degrading	Slow degrading	Steady
1	h_t = mild	h_t = mild	h_t = mild
2	The number of times the health state 'mild' appears in a trajectory	The number of times the health state 'mild' appears in a trajectory	The number of times the health state 'mild' appears in a trajectory
3	h_t = severe	h_t = severe	The number of times the health state 'moderate appears in a trajectory
4	The number of times the health state 'severe appears in a trajectory	The number of times the health state 'severe appears in a trajectory	The number of times the health state 'severe appears in a trajectory
5	h_t*m_t = mild * treatment 1	h_t = moderate	h_t*m_t = mild * treatment 1
6	h_t*m_t = severe * treatment 1	h_t*m_t = mild * treatment 1	h_t = moderate
7	h_{t-5} = mild	h_t*m_t = mild * treatment 2	h_t = severe
8	h_t*m_t = moderate * treatment 2	h_t*m_t = mild * treatment 3	h_t*m_t = moderate * treatment 1
9	h_t = moderate	h_{t-5} = mild	h_t*m_t = severe * treatment 1
10	h_{t-5} = severe	h_t*m_t = severe * treatment 2	h_t*m_t = mild * treatment 2

h_t refers to the health state at time t; h_t = mild represents a binary variable which is 1 if h_t = mild and 0 otherwise.
h_t*m_t refers to an interaction between health state and treatment state at time t; h_t*m_t = mild * treatment 1 represents a binary variable which is 1 if h_t = mild and m_t = treatment 1 and 0 otherwise.

Appendix B Table 2. The top 10 most important features of the final FQI-Ridge model in setup 2

Rank	Fast degrading	Slow degrading	Steady
1	h_t = mild	h_t = mild	h_t = mild
2	The number of times the health state 'mild' appears in a trajectory	h_t = severe	The number of times the health state 'mild' appears in a trajectory
3	h_t = severe	The number of times the health state 'mild' appears in a trajectory	h_t*m_t = mild * treatment 1
4	The number of times the health state 'severe appears in a trajectory	The number of times the health state 'severe appears in a trajectory	The number of times the health state 'moderate' appears in a trajectory
5	h_t*m_t = severe * treatment 2	h_t*m_t = mild * treatment 3	h_t = moderate
6	h_t*m_t = mild * treatment 2	h_t = moderate	h_t = severe
7	h_{t-5} = mild	h_t*m_t = mild * treatment 1	The number of times the health state 'severe appears in a trajectory
8	h_t*m_t = mild * treatment 1	h_{t-5} = mild	h_t*m_t = moderate * treatment 1
9	h_t*m_t = mild * treatment 3	h_t*m_t = mild * treatment 2	h_t*m_t = severe * treatment 1
10	h_{t-5} = severe	h_t*m_t = severe * treatment 2	m_t = treatment 1

h_t refers to the health state at time t; h_t = mild represents a binary variable which is 1 if h_t = mild and 0 otherwise.
m_t refers to the treatment state at time t; m_t = treatment 1 represents a binary variable which is 1 if m_t = treatment 1 and 0 otherwise.
h_t*m_t refers to an interaction between health state and treatment state at time t; h_t*m_t = mild * treatment 1 represents a binary variable which is 1 if h_t = mild and m_t = treatment 1 and 0 otherwise.

b. FQI-RF

A random forest is an ensemble model that constructs multiple decision trees as parallel estimators and outputs the target prediction using the average of predictions from decision trees for a regression problem or using the mode of classes for a classification problem. Here, an ensemble model refers to a model that combines the predictions from several models or a model that consists of multiple models. To better understand a random forest, we provide a brief overview of decision trees.

A decision tree is a model that predicts the interested target by learning decision rules from the data features. The structure of the model has an upside-down tree-like structure that consists of the root node at the top, decision nodes, leaf/terminal nodes, and branches/split. The root node is the first decision node of a decision tree. At each decision node, it contains a condition/rule based on a certain feature on how data should be split, which is a binary decision. A decision parent node can be split into one or more children decision nodes depending on the data. Once the data can no longer be split or the depth of decision tree is reached, we reach the end of the split, which is called terminal nodes.

Although decision trees are interpretable, random forests that consists of multiple decision trees are considered black-box models as it takes into account all predictions from each individual tree. Consequently, we cannot provide decision rules derived from our random forest model. However, we provide the hyperparameters controlling the size of our model as shown in Appendix B Table 3 and the top 10 most important features of our model in Appendix B Table 4 and Appendix B Table 5. For random forest regression, the feature importance is measured by Gini Importance. Gini Importance refers to a measurement of each feature importance by calculating the average over all tress on the decrease in the impurity/misclassification of using the feature in splitting data.

Appendix B Table 3. Hyperparameter controlling the size of our random forest regressor

Hyperparameter	Value
Number of trees in the forest	100
Function measuring the quality of a split	Mean squared error
The maximum depth of the tree	15
The minimum number of samples required to split an internal node	10
Minimum number of samples required to be at a leaf node	5
Number of features	102

Appendix B Table 4. The top 10 most important features of the final FQI-RF model in setup 1

Rank	Fast degrading	Slow degrading	Steady
1	The number of times the health state 'mild' appears in a trajectory	h_t = mild	h_t = mild
2	h_t = mild	The number of times the health state 'mild' appears in a trajectory	The number of times the health state 'mild' appears in a trajectory
3	The number of times the health state 'severe appears in a trajectory	h_{t-2} = mild	The number of times the health state 'severe' appears in a trajectory
4	h_t = severe	The number of times the health state 'severe' appears in a trajectory	h_{t-3} = mild
5	The number of times the health state 'moderate appears in a trajectory	h_{t-1} = mild	h_{t-2} = mild
6	h_{t-1} = severe	The number of times the health state 'moderate' appears in a trajectory	h_{t-1} = severe
7	h_t = moderate	h_{t-1} = severe	h_{t-1} = mild
8	h_{t-2} = mild	h_{t-5} = mild	h_{t-2} = severe
9	h_{t-5} = mild	m_t = treatment 3	m_t = treatment 2
10	h_{t-2} = severe	h_{t-5} = moderate	The number of times the health state 'moderate appears in a trajectory

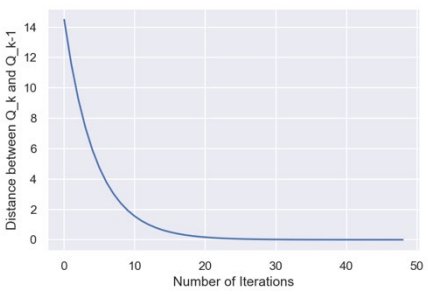
h_t refers to the health state at time t; h_t = mild represents a decision rule on whether h_t = mild or not.
m_t refers to the treatment state at time t; m_t = treatment 2 represents a decision rule on whether m_t = treatment 2 or not.
h_t*m_t refers to an interaction between health state and treatment state at time t; h_t*m_t = mild * treatment 1 represents a decision rule on whether h_t = mild and m_t = treatment 1 or not.

Appendix B Table 5. The top 10 most important features of the final FQI-RF model in setup 2

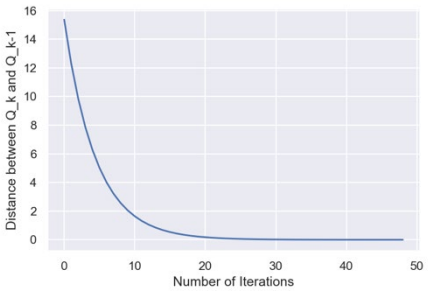
Rank	Fast degrading	Slow degrading	Steady
1	The number of times the health state 'mild' appears in a trajectory	h_t = mild	h_t = mild
2	h_t = mild	The number of times the health state 'mild' appears in a trajectory	The number of times the health state 'mild' appears in a trajectory
3	h_t = severe	The number of times the health state 'severe' appears in a trajectory	The number of times the health state 'severe' appears in a trajectory
4	The number of times the health state 'severe' appears in a trajectory	h_{t-1} = mild	h_{t-3} = mild
5	The number of times the health state 'moderate' appears in a trajectory	h_{t-2} = mild	h_{t-2} = mild
6	h_{t-1} = severe	The number of times the health state 'moderate' appears in a trajectory	h_{t-1} = mild
7	h_t = moderate	h_t = moderate	h_{t-1} = severe
8	h_{t-5} = mild	h_{t-5} = mild	h_{t-2} = severe
9	h_{t-2} = severe	h_{t-1} = severe	The number of times the health state 'moderate' appears in a trajectory
10	h_{t-1} = mild	h_{t-2} = severe	h_{t-4} = mild

h_t refers to the health state at time t; h_t = mild represents a decision rule on whether h_t = mild or not.
m_t refers to the treatment state at time t; m_t = treatment 2 represents a decision rule on whether m_t = treatment 2 or not.
h_t*m_t refers to an interaction between health state and treatment state at time t; h_t*m_t = mild * treatment 1 represents a decision rule on whether h_t = mild and m_t = treatment 1 or not.

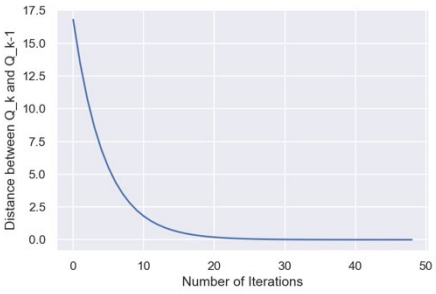
2. Additional results for simulated experiments



Fast-degrading progression

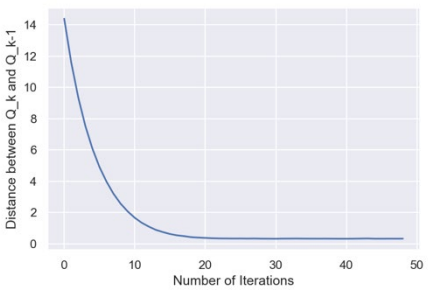


Slow-degrading progression

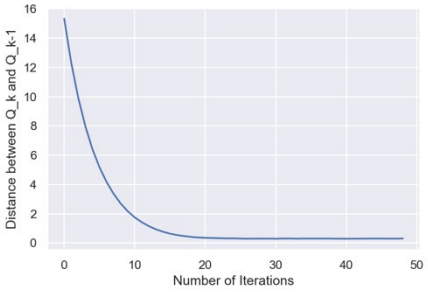


Steady progression

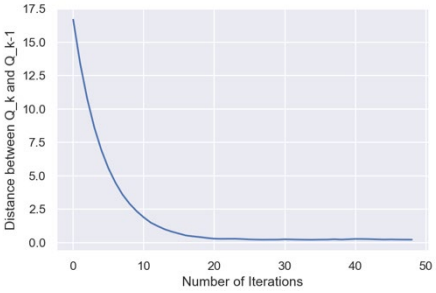
Appendix B Figure 1. Convergence of the Q-function approximated by FQI-Ridge in the base case by progression group in setup 1



Fast-degrading progression

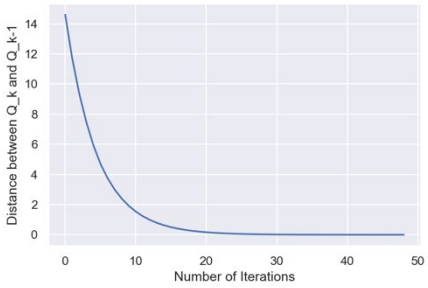


Slow-degrading progression

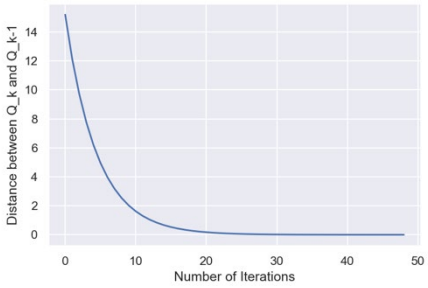


Steady progression

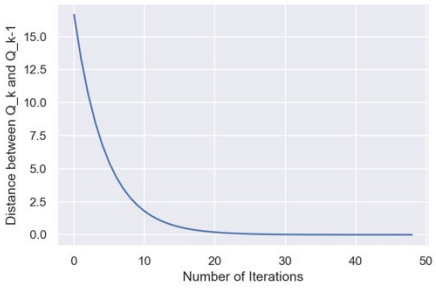
Appendix B Figure 2. Convergence of the Q-function approximated by FQI-RF in the base case by progression group in setup 1



Fast-degrading progression

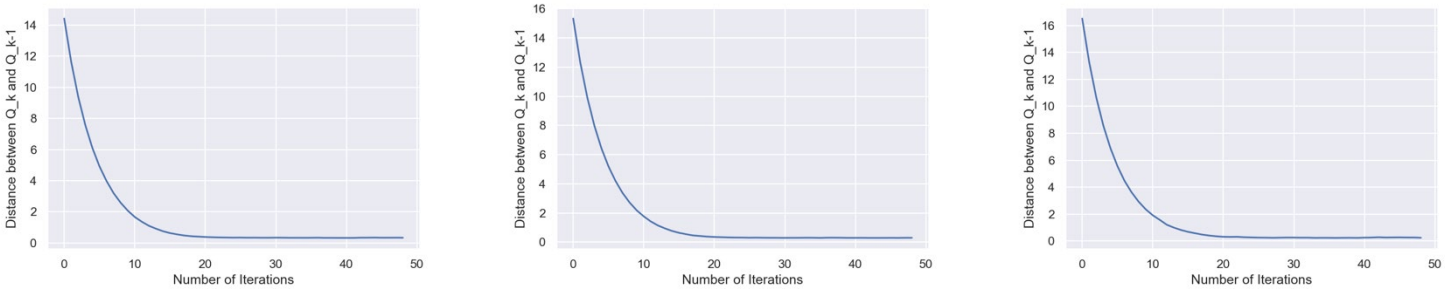


Slow-degrading progression



Steady progression

Appendix B Figure 3. Convergence of the Q-function approximated by FQI-Ridge in the base case by progression group in setup 2



Fast-degrading progression | Slow-degrading progression | Steady progression

Appendix B Figure 4. Convergence of the Q-function approximated by FQI-RF in the base case by progression group in setup 2

References

1 Weill JA, Stigler M, Deschenes O, Springborn MR. **Social distancing responses to COVID-19 emergency declarations strongly differentiated by income**. *Proc Natl Acad Sci USA* 2020; **117**:19658–19660.

2 Keirns CC, Goold, Susan Dorr. **Patient-Centered Care and Preference-Sensitive Decision Making**. *JAMA* 2009; **302**:1805.

3 Marshall DA, Gonzalez JM, MacDonald KV, Johnson FR. **Estimating Preferences for Complex Health Technologies: Lessons Learned and Implications for Personalized Medicine**. *Value in Health* 2017; **20**:32–39.

4 UNAIDS. AIDsinfo Global data on HIV epidemiology and response. https://aidsinfo.unaids.org/ (accessed 20 Jun2021).

5 Boerma RS, Sigaloff KCE, Akanmu AS, Inzaule S, Boele van Hensbroek M, Rinke de Wit TF, *et al.* **Alarming increase in pretreatment HIV drug resistance in children living in sub-Saharan Africa: a systematic review and meta-analysis**. *J Antimicrob Chemother* 2017; **72**:365–371.

6 WHO. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach. Geneva: World Health Organization; 2013. https://apps.who.int/iris/bitstream/handle/10665/85321/9789241505727_eng.pdf;jsessionid=E999278C23AF0E18688D10A2612F3999?sequence=1 (accessed 23 Dec2019).

7 Kityo C, Boerma RS, Sigaloff KCE, Kaudha E, Calis JCJ, Musiime V, *et al.* **Pretreatment HIV drug resistance results in virological failure and accumulation of additional resistance mutations in Ugandan children**. *Journal of Antimicrobial Chemotherapy* 2017; **72**:2587–2595.

8 Hamers RL, Schuurman R, Sigaloff KC, Wallis CL, Kityo C, Siwale M, *et al.* **Effect of pretreatment HIV-1 drug resistance on immunological, virological, and drug-resistance outcomes of first-line antiretroviral treatment in sub-Saharan Africa: a multicentre cohort study**. *The Lancet Infectious Diseases* 2012; **12**:307–317.

9 Adedimeji A, Edmonds A, Hoover D, Shi Q, Sinayobye J d'Amour, Nduwimana M, *et al.* **Characteristics of HIV-Infected Children at Enrollment into Care and at Antiretroviral Therapy Initiation in Central Africa**. *PLoS ONE* 2017; **12**:e0169871.

10 Davies M-A, Phiri S, Wood R, Wellington M, Cox V, Bolton-Moore C, *et al.* **Temporal Trends in the Characteristics of Children at Antiretroviral Therapy Initiation in Southern Africa: The IeDEA-SA Collaboration**. *PLoS ONE* 2013; **8**:e81037.

11 WHO. Update of recommendations on first- and second-line antiretroviral regimens. Geneva, Switzerland: World Health Organization; 2019.

12 Llibre JM, Hung C-C, Brinson C, Castelli F, Girard P-M, Kahl LP, *et al.* **Efficacy, safety, and tolerability of dolutegravir-rilpivirine for the maintenance of virological suppression in adults with HIV-1: phase 3, randomised, non-inferiority SWORD-1 and SWORD-2 studies**. *The Lancet* 2018; **391**:839–849.

13 Viani RM, Alvero C, Fenton T, Acosta EP, Hazra R, Townley E, *et al.* **Safety, Pharmacokinetics and Efficacy of Dolutegravir in Treatment-experienced HIV-1 Infected Adolescents: Forty-eight-week Results from IMPAACT P1093**. *The Pediatric Infectious Disease Journal* 2015; **34**:1207–1213.

14 Bruzzese E, Lo Vecchio A, Smarrazzo A, Tambaro O, Palmiero G, Bonadies G, *et al.* **Dolutegravir-based anti-retroviral therapy is effective and safe in HIV–infected paediatric patients**. *Ital J Pediatr* 2018; **44**:37.

15 WHO. Transition to new antiretroviral drugs in HIV programmes: clinical and programmatic considerations. Geneva: World Health Organization; 2017.

16 WHO. New high-quality antiretroviral therapy to be launched in South Africa, Kenya and over 90 low- and middle-income countries at reduced price. 2017.https://www.who.int/hiv/mediacentre/news/high-quality-arv-reduced-price/en/ (accessed 20 Aug2020).

17 Duarte HA, Babigumira JB, Enns EA, Stauffer DC, Shafer RW, Beck IA, *et al.* **Cost-effectiveness analysis of pre-ART HIV drug resistance testing in Kenyan women**. *EClinicalMedicine* 2020; **22**:100355.

18 Nichols BE, Sigaloff KC, Kityo C, Hamers RL, Baltussen R, Bertagnolio S, *et al.* **Increasing the use of second-line therapy is a cost-effective approach to prevent the spread of drug-resistant HIV: a mathematical modelling study**. *Journal of the International AIDS Society* 2014; **17**:19164.

19 Desmonde S, Eboua FT, Malateste K, Dicko F, Ekouévi DK, Ngbeché S, *et al.* **Determinants of durability of first-line antiretroviral therapy regimen and time from first-line failure to second-line antiretroviral therapy initiation:** *AIDS* 2015; **29**:1527–1536.

20 Wools-Kaloustian K, Marete I, Ayaya S, Sohn AH, Van Nguyen L, Li S, *et al.* **Time to First-Line ART Failure and Time to Second-Line ART Switch in the IeDEA Pediatric Cohort:** *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2018; **78**:221–230.

21 WHO. *Consolidated Guidelines on HIV Prevention, Diagnosis, Treatment and Care for Key Populations*. World Health Organization; 2016. http://proxy.library.carleton.ca/loginurl=https://www.deslibris.ca/ID/10063272 (accessed 22 Jul2020).

22 Ciaranello AL, Morris BL, Walensky RP, Weinstein MC, Ayaya S, Doherty K, *et al.* **Validation and Calibration of a Computer Simulation Model of Pediatric HIV Infection**. *PLoS ONE* 2013; **8**:e83389.

23 Marston M, Zaba B, Salomon JA, Brahmbhatt H, Bagenda D. **Estimating the Net Effect of HIV on Child Mortality in African Populations Affected by Generalized HIV Epidemics:** *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2005; **38**:219–227.

24 Violari A, Paed FC, Lindsey JC, Hughes MD, Mujuru HA, Chi BH, *et al.* **Nevirapine versus Ritonavir-Boosted Lopinavir for HIV-Infected Children**. *n engl j med* 2012; :10.

25 Ledergerber B. **Predictors of trend in CD4-positive T-cell count and mortality among HIV-1-infected individuals with virological failure to all three antiretroviral-drug classes**. *The Lancet* 2004; **364**:51–62.

26 Boerma RS, Bunupuradah T, Dow D, Fokam J, Kariminia A, Lehman D, *et al.* **Multicentre analysis of second-line antiretroviral treatment in HIV-infected children: adolescents at high risk of failure**. *Journal of the International AIDS Society* 2017; **20**:21930.

27 Carlucci JG, Liu Y, Friedman H, Pelayo BE, Robelin K, Sheldon EK, *et al.* **Attrition of HIV-exposed infants from early infant diagnosis services in low- and middle-income countries: a systematic review and meta-analysis**. *J Intern AIDS Soc* 2018; **21**:e25209.

28 Boerma RS, Boender TS, Bussink AP, Calis JCJ, Bertagnolio S, Rinke de Wit TF, *et al.* **Suboptimal Viral Suppression Rates Among HIV-Infected Children in Low- and Middle-Income Countries: A Meta-analysis**. *Clin Infect Dis* 2016; **63**:1645–1654.

29 WHO-CHOICE. Estimates of Unit Costs for Patient Services for South Africa. https://www.who.int/choice/country/zaf/cost/en/ (accessed 20 Sep2020).

30 Institute for Health Metrics and Evaluation (IHME). Health Service Provision in Kenya: Assessing Facility Capacity, Costs of Care, and Patient Perspectives. Seattle, WA: IHME; 2014.

31 Institute for Health Metrics and Evaluation (IHME). Health Service Provision in Ghana: Assessing Facility Capacity, Costs of Care, and Patient Perspectives. Seattle, WA: IHME; 2015.

32 Institute for Health Metrics and Evaluation (IHME). Health Service Provision in Uganda: Assessing Facility Capacity, Costs of Care, and Patient Perspectives. Seattle, WA: IHME; 2014.

33 Institute for Health Metrics and Evaluation (IHME). Health Service Provision in Zambia: Assessing Facility Capacity, Costs of Care, and Patient Perspectives. Seattle, WA: IHME; 2014.

34 World Health Organization Global Health Expenditure. World Bank Open Data. 2016.https://data.worldbank.org/ (accessed 29 Dec2020).

35 Ciaranello AL, Doherty K, Penazzato M, Lindsey JC, Harrison L, Kelly K, *et al.* **Cost-effectiveness of first-line antiretroviral therapy for HIV-infected African children less than 3 years of age:** *AIDS* 2015; **29**:1247–1259.

36 Desmonde S, Frank SC, Coovadia A, Dahourou DL, Hou T, Abrams EJ, *et al.* **Cost-Effectiveness of Preemptive Switching to Efavirenz-Based Antiretroviral Therapy for Children With Human Immunodeficiency Virus**. *Open Forum Infectious Diseases* 2019; **6**:ofz276.

37 Phillips AN, Cambiano V, Nakagawa F, Revill P, Jordan MR, Hallett TB, *et al.* **Cost-effectiveness of public-health policy options in the presence of pretreatment NNRTI drug resistance in sub-Saharan Africa: a modelling study**. *The Lancet HIV* 2018; **5**:e146–e154.

38 Ochalek J, Lomas J, Claxton K. **Estimating health opportunity costs in low-income and middle-income countries: a novel approach and evidence from cross-country data**. *BMJ Glob Health* 2018; **3**:e000964.

39 Woods B, Revill P, Sculpher M, Claxton K. **Country-Level Cost-Effectiveness Thresholds: Initial Estimates and the Need for Further Research**. *Value in Health* 2016; **19**:929–935.

40 Horton S, Gelband H, Jamison D, Levin C, Nugent R, Watkins D. **Ranking 93 health interventions for low- and middle-income countries by cost-effectiveness**. *PLoS ONE* 2017; **12**:e0182951.

41 Cruciani M, Parisi SG. **Dolutegravir based antiretroviral therapy compared to other combined antiretroviral regimens for the treatment of HIV-infected naive patients: A systematic review and meta-analysis**. *PLoS ONE* 2019; **14**:e0222229.

42 Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017 (GBD 2017) Burden by Risk 1990-2017. Seattle, United States: Institute for Health Metrics and Evaluation (IHME); 2018. http://ghdx.healthdata.org/gbd-2017/data-input-sources (accessed 29 Mar2020).

43 Greenberg PE, Fournier A-A, Sisitsky T, Pike CT, Kessler RC. **The Economic Burden of Adults With Major Depressive Disorder in the United States (2005 and 2010)**. *J Clin Psychiatry* 2015; **76**:155–162.

44 Brody DJ, Gu Q. Antidepressant Use Among Adults: United States, 2015–2018. Hyattsville, MD: National Center for Health Statistics; 2020.

45 American Psychological Association. *Practice Guideline for the Treatment of Patients With Major Depressive Disorder*. 3rd ed. American Psychological Association; 2010. https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd.pdf (accessed 24 Feb2021).

46 Department of Veterans Affairs and Department of Defense. VA/DoD Clinical Practice Guideline for the Management of Major Depressive Disorder. Version 3.0. Department of Veterans Affairs and Department of Defense; 2016.

47 American Psychological Association. Clinical Practice Guideline for the Treatment of Depression Across Three Age Cohorts American Psychological Association Guideline Development Panel for the Treatment of Depressive Disorders. American Psychological Association; 2019. https://www.apa.org/depression-guideline/guideline.pdf (accessed 1 Dec2019).

48 Trangle M, Gursky J, Haight R, Hardwig J, Hinnenkamp T, Kessler D, *et al.* Adult Depression in Primary Care. Institute for Clinical Systems Improvement; 2016.

49 Simon GE, Perlis RH. **Personalized Medicine for Depression: Can We Match Patients With Treatments?** *Am J Psychiatry* 2010; :11.

50 Cuijpers P, Reynolds CF, Donker T, Li J, Andersson G, Beekman A. **PERSONALIZED TREATMENT OF ADULT DEPRESSION: MEDICATION, PSYCHOTHERAPY, OR BOTH? A SYSTEMATIC REVIEW. Depress Anxiety**. *Depress Anxiety* 2012; **29**:855–864.

51 Jameson JL. **Precision Medicine — Personalized, Problematic, and Promising**. *n engl j med* 2015; :6.

52 Kosorok MR, Laber EB. **Precision Medicine**. 2019; :26.

53 Ginsburg GS, Phillips KA. **Precision Medicine: From Science To Value**. *Health Affairs* 2018; **37**:694–701.

54 Shechter SM, Bailey MD, Schaefer AJ, Roberts MS. **The Optimal Time to Initiate HIV Therapy Under Ordered Health States**. *Operations Research* 2008; **56**:20–33.

55 Kurt M, Denton BT, Schaefer AJ, Shah ND, Smith SA. **The structure of optimal statin initiation policies for patients with Type 2 diabetes**. *IIE Transactions on Healthcare Systems Engineering* 2011; **1**:49–65.

56 Denton BT, Kurt M, Shah ND, Bryant SC, Smith SA. **Optimizing the Start Time of Statin Therapy for Patients with Diabetes**. *Med Decis Making* 2009; **29**:351–367.

57 Lee CP, Chertow GM, Zenios SA. **Optimal Initiation and Management of Dialysis Therapy**. *Operations Research* 2008; **56**:1428–1449.

58 Erenay FS, Alagoz O, Said A. **Optimizing Colonoscopy Screening for Colorectal Cancer Prevention and Surveillance**. *M&SOM* 2014; **16**:381–400.

59 Steimle LN, Denton BT. Markov decision processes for screening and treatment of chronic diseases. In: *Markov Decision Processes in Practice*.Cham: Springer; 2017. pp. 189–222.

60 Chen X, Shachter RD, Kurian AW, Rubin DL. **Dynamic strategy for personalized medicine: An application to metastatic breast cancer**. *Journal of Biomedical Informatics* 2017; **68**:50–57.

61 Schell GJ, Marrero WJ, Lavieri MS, Sussman JB, Hayward RA. **Data-Driven Markov Decision Process Approximations for Personalized Hypertension Treatment Planning**. *MDM Policy & Practice* 2016; **1**:238146831667421.

62 Bellman R. **Dynamic Programming**. *Science* 1996; **153**:34–37.

63 Sutton RS, Barto AG. *Reinforcement learning: An introduction*. MIT press; 2018.

64 Schulte PJ, Tsiatis AA, Laber EB, Davidian M. **Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes**. *Statist Sci* 2014; **29**:640–661.

65 Yu C, Dong Y, Liu J, Ren G. **Incorporating causal factors into reinforcement learning for dynamic treatment regimes in HIV**. *BMC Med Inform Decis Mak* 2019; **19**:60.

66 Zhang B, Tsiatis AA, Laber EB, Davidian M. **A Robust Method for Estimating Optimal Treatment Regimes**. *Biometrics* 2012; **68**:1010–1018.

67 Zhao Y, Zeng D, Socinski MA, Kosorok MR. **Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer**. *Biometrics* 2011; **67**:1422–1433.

68 Jalalimanesh A, Shahabi Haghighi H, Ahmadi A, Soltani M. **Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning**. *Mathematics and Computers in Simulation* 2017; **133**:235–248.

69 Ernst D, Geurts P, Wehenkel L. **Tree-Based Batch Mode Reinforcement Learning**. *Journal of Machine Learning Research* 2005; **6**:503–556.

70 Guez A, Vincent RD, Avoli M, Pineau J. **Adaptive Treatment of Epilepsy via Batch-mode Reinforcement Learning**. *AAAI* 2008; :1671–1678.

71 Myhre JN, Launonen IK, Wei S, Godtliebsen F. CONTROLLING BLOOD GLUCOSE LEVELS IN PATIENTS WITH TYPE 1 DIABETES USING FITTED Q-ITERATIONS AND FUNCTIONAL FEATURES. In: *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*.Aalborg: IEEE; 2018. pp. 1–6.

72 Escandell-Montero P, Chermisi M, Martínez-Martínez JM, Gómez-Sanchis J, Barbieri C, Soria-Olivas E, *et al.* **Optimization of anemia treatment in hemodialysis patients via reinforcement learning**. *Artificial Intelligence in Medicine* 2014; **62**:47–60.

73 Zhao Y-Q, Laber EB. **Estimation of optimal dynamic treatment regimes**. *Clinical Trials* 2014; **11**:400–407.

74 Laber EB, Linn KA, Stefanski LA. **Interactive model building for Q-learning**. *Biometrika* 2014; **101**:831–847.

75 Cheung YK, Chakraborty B, Davidson KW. **Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program: SMART with Adaptive Randomization**. *Biom* 2015; **71**:450–459.

76 Chakraborty B, Ghosh P, Moodie EEM, Rush AJ. **Estimating optimal shared-parameter dynamic regimens with application to a multistage depression clinical trial: Shared-Parameter Dynamic Regimens**. *Biom* 2016; **72**:865–876.

77 Shortreed SM, Laber E, Lizotte DJ, Stroup TS, Pineau J, Murphy SA. **Informing sequential clinical decision-making through reinforcement learning: an empirical study**. *Mach Learn* 2011; **84**:109–136.

78 Ernst D, Stan G-B, Goncalves J, Wehenkel L. Clinical data based optimal STI strategies for HIV: a reinforcement learning approach. In: *Proceedings of the 45th IEEE Conference on Decision and Control*.San Diego, CA: IEEE; 2006. pp. 667–672.

79 Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, *et al.* **Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design**. *Controlled Clinical Trials* 2004; **25**:119–142.

80 Linn KA, Laber EB, Stefanski LA. **Interactive Q-Learning for Quantiles**. *Journal of the American Statistical Association* 2017; **112**:638–649.

81 Kroenke K, Spitzer RL, Williams JBW. **The PHQ-9: Validity of a brief depression severity measure**. *J Gen Intern Med* 2001; **16**:606–613.

82 Ross EL, Vijan S, Miller EM, Valenstein M, Zivin K. **The Cost-Effectiveness of Cognitive Behavioral Therapy Versus Second-Generation Antidepressants for Initial Treatment of Major Depressive Disorder in the United States: A Decision Analytic Model**. *Ann Intern Med* 2019; **171**:785.

83 Cuijpers P, Vogelzangs N, Twisk J, Kleiboer A, Li J, Penninx BW. **Comprehensive Meta-Analysis of Excess Mortality in Depression in the General Community Versus Patients With Specific Illnesses**. *AJP* 2014; **171**:453–462.

84 Arias E, Xu J. United States Life Tables, 2017. Hyattsville, MD: National Center for Health Statistics; 2019.

85 Kolovos S, Bosmans JE, van Dongen JM, van Esveld B, Magai D, van Straten A, *et al.* **Utility scores for different health states related to depression: individual participant data analysis**. *Qual Life Res* 2017; **26**:1649–1658.

86 Ont Health Technol Assess Ser [Internet]. **Internet-delivered cognitive behavioural therapy for major depression and anxiety disorders: a health technology assessment**. 2019; **19**:1–199.

87 Nguyen K-H, Gordon LG. **Cost-Effectiveness of Repetitive Transcranial Magnetic Stimulation versus Antidepressant Therapy for Treatment-Resistant Depression**. *Value in Health* 2015; **18**:597–604.

88 Thase ME, Friedman ES, Biggs MM, Wisniewski SR, Trivedi MH, Luther JF, *et al.* **Cognitive Therapy Versus Medication in Augmentation and Switch Strategies as Second-Step Treatments: A STAR*D Report**. *Am J Psychiatry* 2007; :14.

89 Lin Y, Huang S, Simon GE, Liu S. **Cost-effectiveness analysis of prognostic-based depression monitoring**. *IISE Transactions on Healthcare Systems Engineering* 2019; **9**:41–54.

90 Simon GE, Yarborough BJ. **Good News: Artificial Intelligence in Psychiatry Is Actually Neither**. *PS* 2020; **71**:219–220.

91 Beam AL, Kohane IS. **Big Data and Machine Learning in Health Care**. *JAMA* 2018; **319**:1317.

92 Twisk J, Hoekstra T. **Classifying developmental trajectories over time should be done with great caution: a comparison between methods**. *Journal of Clinical Epidemiology* 2012; **65**:1078–1087.

93 van Lang NDJ, Ferdinand RF, Ormel J, Verhulst FC. **Latent class analysis of anxiety and depressive symptoms of the Youth Self-Report in a general population sample of young adolescents**. *Behaviour Research and Therapy* 2006; **44**:849–860.

94 Connell AM, Frye AA. **Growth mixture modelling in developmental psychology: overview and demonstration of heterogeneity in developmental trajectories of adolescent antisocial behaviour**. *Inf Child Develop* 2006; **15**:609–621.

95 Lin Y, Huang S, Simon GE, Liu S. **Analysis of depression trajectory patterns using collaborative learning**. *Mathematical Biosciences* 2016; **282**:191–203.

96 Craig BA, Sendi PP. **Estimation of the transition matrix of a discrete-time Markov chain**. *Health Econ* 2002; **11**:33–42.

97 Wongvibulsin S, Martin SS, Saria S, Zeger SL, Murphy SA. **An Individualized, Data-Driven Digital Approach for Precision Behavior Change**. *American Journal of Lifestyle Medicine* 2020; **14**:289–293.

98 Jaimes L, Llofriu M, Raij A. A Stress-Free Life: Just-in-Time Interventions for Stress via Real-Time Forecasting and Intervention Adaptation. In: *Proceedings of the 9th International Conference on Body Area Networks*.London, Great Britain: ICST; 2014. doi:10.4108/icst.bodynets.2014.258237

99 Liao P, Greenewald K, Klasnja P, Murphy S. **Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity**. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2020; **4**:1–22.

100 Rosen CS, Morland LA, Glassman LH, Marx BP, Weaver K, Smith CA, *et al.* **Virtual mental health care in the Veterans Health Administration's immediate response to coronavirus disease-19.** *American Psychologist* 2021; **76**:26–38.

101 Ontario health technology. **Ontario health technology assessment series: Internet-delivered cognitive behavioural therapy for major depression and anxiety disorders: A health technology assessment**. *Ontario Health Technology Assessment Series* 2019; **19**.

102 Hubley S, Lynch SB, Schneck C, Thomas M, Shore J. **Review of key telepsychiatry outcomes**. *WJP* 2016; **6**:269.

103 Chan IS, Ginsburg GS. **Personalized Medicine: Progress and Promise**. *Annu Rev Genom Hum Genet* 2011; **12**:217–244.

104 Goetz LH, Schork NJ. **Personalized medicine: motivation, challenges, and progress**. *Fertility and Sterility* 2018; **109**:952–963.

105 Preference Collaborative Review Group. **Patients' preferences within randomised trials: systematic review and patient level meta-analysis**. *BMJ* 2008; **337**:a1864–a1864.

106 Weernink MGM, Janus SIM, van Til JA, Raisch DW, van Manen JG, IJzerman MJ. **A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making**. *Pharm Med* 2014; **28**:175–185.

107 Carson RT, Louviere JJ. **A Common Nomenclature for Stated Preference Elicitation Approaches**. *Environ Resource Econ* 2011; **49**:539–559.

108 Lizotte DJ, Bowling M, Murphy SA. **Linear Fitted-Q Iteration with Multiple Reward Functions**. *The Journal of Machine Learning Research* 2012; **13**:3253–3295.

109 Butler EL, Laber EB, Davis SM, Kosorok MR. **Incorporating Patient Preferences into Estimation of Optimal Individualized Treatment Rules: Incorporating Patient Preferences for Optimal Treatment**. *Biom* 2018; **74**:18–26.

110 Asoh H, Shiro M, Akaho S, Kamishima T, Hasida K, Aramaki E, *et al.* An Application of Inverse Reinforcement Learning to Medical Records of Diabetes Treatment. ; 2013.

111 Erkin Z, Bailey MD, Maillart LM, Schaefer AJ, Roberts MS. **Eliciting Patients' Revealed Preferences: An Inverse Markov Decision Process Approach**. *Decision Analysis* 2010; **7**:358–365.

112 Ng AY, Russell S. Algorithms for inverse reinforcement learning. In: *17th International Conference on Machine Learning*.; 2000. pp. 663–670.

113 Abbeel P, Ng AY. Apprenticeship learning via inverse reinforcement learning. In: *Twenty-first international conference on Machine learning - ICML '04*.Banff, Alberta, Canada: ACM Press; 2004. p. 1.

114 Ratliff ND, Bagnell JA, Zinkevich MA. Maximum margin planning. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*.Pittsburgh, Pennsylvania: ACM Press; 2006. pp. 729–736.

115 Ziebart BD, Maas A, Bagnell JA, Dey AK. Maximum Entropy Inverse Reinforcement Learning. In: *The 23rd AAAI Conference on Artificial Intelligence*.; 2008. pp. 1433–1438.

116 Ikenaga A, Arai S. Inverse Reinforcement Learning Approach for Elicitation of Preferences in Multi-objective Sequential Optimization. In: *2018 IEEE International Conference on Agents (ICA)*.Singapore: IEEE; 2018. pp. 117–118.

117 Grollman DH, Billard A. Donut as I do: Learning from failed demonstrations. In: *2011 IEEE International Conference on Robotics and Automation*.Shanghai, China: IEEE; 2011. pp. 3804–3809.

118 Shiarlis K, Messias J, Whiteson S. Inverse Reinforcement Learning from Failure. ; 2016. pp. 1060–1068.

119 Brown DS, Goo W, Nagarajan P, Niekum S. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations. In: *International conference on machine learning*.; 2019. pp. 783–792.

120 Chen L, Pu P. Survey of Preference Elicitation Methods. EPFL; 2004.

121 Rabbi M, Aung MH, Zhang M, Choudhury T. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*.Osaka, Japan: ACM Press; 2015. pp. 707–718.

122 Kekitiinwa A, Lee KJ, Walker AS, Maganda A, Doerholt K, Kitaka SB, *et al.* **Differences in Factors Associated With Initial Growth, CD4, and Viral Load Responses to ART in HIV-Infected Children in Kampala, Uganda, and the United Kingdom/Ireland:** *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2008; **49**:384–392.

123 PENPACT-1 (PENTA 9/PACTG 390) Study Team. **First-line antiretroviral therapy with a protease inhibitor versus non-nucleoside reverse transcriptase inhibitor and switch at higher versus low viral load in HIV-infected children: an open-label, randomised phase 2/3 trial**. *The Lancet Infectious Diseases* 2011; **11**:273–283.

124 Adjé-Touré C, Hanson DL, Talla-Nzussouo N, Borget M-Y, Kouadio LY, Tossou O, *et al.* **Virologic and Immunologic Response to Antiretroviral Therapy and Predictors of HIV Type 1 Drug Resistance in Children Receiving Treatment in Abidjan, Côte d'Ivoire**. *AIDS Research and Human Retroviruses* 2008; **24**:911–917.

125 Barry O, Powell J, Renner L, Bonney EY, Prin M, Ampofo W, *et al.* **Effectiveness of first-line antiretroviral therapy and correlates of longitudinal changes in CD4 and viral load among HIV-infected children in Ghana**. *BMC Infect Dis* 2013; **13**:476.

126 Han J, Mu W, Zhao H, Hao Y, Song C, Zhou H, *et al.* **HIV-1 low-level viremia affects T cell activation rather than T cell development in school-age children, adolescents, and**

young adults during antiretroviral therapy. *International Journal of Infectious Diseases* 2020; **91**:210–217.

127 Desmonde S, Neilan AM, Musick B, Patten G, Chokephaibulkit K, Edmonds A, *et al.* **Time-varying age- and CD4-stratified rates of mortality and WHO stage 3 and stage 4 events in children, adolescents and youth 0 to 24 years living with perinatally acquired HIV, before and after antiretroviral therapy initiation in the paediatric IeDEA Global Cohort Consortium**. *J Intern AIDS Soc* 2020; **23**. doi:10.1002/jia2.25617

128 Vanni T, Karnon J, Madan J, White RG, Edmunds WJ, Foss AM, *et al.* **Calibrating Models in Economic Evaluation: A Seven-Step Approach**. *PharmacoEconomics* 2011; **29**:35–49.

129 Violari A, Lindsey JC, Hughes MD, Mujuru HA, Barlow-Mosha L, Kamthunzi P, *et al.* **Nevirapine versus Ritonavir-Boosted Lopinavir for HIV-Infected Children**. *N Engl J Med* 2012; **366**:2380–2389.

130 Sutcliffe CG, Bolton-Moore C, van Dijk JH, Cotham M, Tambatamba B, Moss WJ. **Secular trends in pediatric antiretroviral treatment programs in rural and urban Zambia: a retrospective cohort study**. *BMC Pediatr* 2010; **10**:54.

131 Fatti G, Bock P, Eley B, Mothibi E, Grimwood A. **Temporal Trends in Baseline Characteristics and Treatment Outcomes of Children Starting Antiretroviral Treatment: An Analysis in Four Provinces in South Africa, 2004–2009**. *J Acquir Immune Defic Syndr* 2011; **58**:8.

132 Davies M-A, Phiri S, Wood R, Wellington M, Cox V, Bolton-Moore C, *et al.* **Temporal Trends in the Characteristics of Children at Antiretroviral Therapy Initiation in Southern Africa: The IeDEA-SA Collaboration**. *PLoS ONE* 2013; **8**:e81037.

133 Muenchhoff M, Adland E, Roider J, Kløverpris H, Leslie A, Boehm S, *et al.* **Differential Pathogen-Specific Immune Reconstitution in Antiretroviral Therapy-Treated Human Immunodeficiency Virus-Infected Children**. *The Journal of Infectious Diseases* 2019; **219**:1407–1417.

134 Prendergast AJ, Szubert AJ, Berejena C, Pimundu G, Pala P, Shonhai A, *et al.* **Baseline Inflammatory Biomarkers Identify Subgroups of HIV-Infected African Children With Differing Responses to Antiretroviral Therapy**. *J Infect Dis* 2016; **214**:226–236.

135 Panel on Antiretroviral Therapy and Medical Management of Children Living with HIV. Guidelines for the Use of Antiretroviral Agents in Pediatric HIV Infection. https://aidsinfo.nih.gov/contentfiles/lvguidelines/pediatricguidelines.pdf (accessed 12 Aug2020).

136 Chung MH, McGrath CJ, Beck IA, Levine M, Milne RS, So I, *et al.* **Evaluation of the management of pretreatment HIV drug resistance by oligonucleotide ligation assay: a randomised controlled trial**. *The Lancet HIV* 2020; **7**:e104–e112.

137 Dugdale CM, Ciaranello AL, Bekker L-G, Stern ME, Myer L, Wood R, *et al.* **Risks and Benefits of Dolutegravir- and Efavirenz-Based Strategies for South African Women With HIV of Child-Bearing Potential: A Modeling Study**. *Ann Intern Med* 2019; **170**:614.

138 Davies M-A, Moultrie H, Eley B, Rabie H, Van Cutsem G, Giddy J, *et al.* **Virologic Failure and Second-Line Antiretroviral Therapy in Children in South Africa—The IeDEA Southern Africa Collaboration**. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2011; **56**:270–278.

139 Panel on Antiretroviral Therapy and Medical Management of Children Living with HIV. Guidelines for the Use of Antiretroviral Agents in Pediatric HIV. http://aidsinfo.nih.gov/contentfiles/lvguidelines/pediatricguidelines.pdf