Modern Methods in Precision Medicine

Kyra (Jingyi) Gan

April 2022

Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Andrew Li (Co-chair)
Sridhar Tayur (Co-chair)
Tinglong Dai
Zachary Lipton
Alan Scheller-Wolf

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Operations Research.*

Contents

Chapter 1

Introduction

As the concept of precision medicine spreads, there is a growing need for developing better algorithms that a) are sample efficient (i.e., require fewer samples to achieve the same accuracy level), b) think beyond association (to identify the *causation* hidden in the data), and c) provide insights to medical practice. In this dissertation, we investigate various problems in precision medicine, the topics ranging from *opioid use disorder* (OUD) and cancer treatment, to *sickle cell disease* (SCD). We leverage tools from stochastic learning, causal inference, and machine learning, with the objective of reducing healthcare expenditure and improving the quality of care.

One of the US's most recent health crises is the opioid overdose epidemic, and the resource that we have to reverse is epidemic is limited. While various OUD treatments have shown to be effective on a population level, individual patients react differently to these treatments. Wearable devices, on the other hand, can potentially revolutionize treatments for OUD by measuring patient responses to different treatment regimens in real-time, enabling the development of personalized treatments. However, before we deploy the use of wearable devices in OUD treatments, we first need to understand the practicality and the cost-effectiveness of such devices. Thus motivated, in Chapter 2, we evaluate the use of wearable devices in OUD treatments when the budget is limited. In particular, we consider a variety of wearable devices with different features, sensitivities, and costs, and model our problem using a finite-horizon, non-stationary *constrained partially observable Markov decision process* (CPOMDP). To facilitate the solution of our model, we provide a novel budget reformulation that finds all optimal solutions lying on the original formulation's solution's convex hull. Next, we show our reformulation can be solved using a binary search in conjunction with an exact POMDP algorithm. We apply those elements, using extracted transition matrices and rewards from past literature, to perform a numerical study to investigate the value of incorporating different wearables in treatments for OUD under scenarios described by different levels of budget, wearable precision, and patient *treatment adherence* (TA). We find that wearables can be valuable at moderate budgets for patients with low or moderate TA; this benefit increases as the wearable accuracy increases. Outside of these settings, either the marginal benefit of wearables is negligible relative to their cost, or their use increases the patients' risk of overdose to an unacceptable degree.

Chapters 3 and 4 both relate to cancer research. One of the fundamental goals in cancer research is to identify the genetic mutations that can *cause* cancer. If such mutations were identified, then targeted drugs can be produced to block the effect of these mutations and hence curing cancer. Since editing human genome is clinically unsafe at the currently stage, to derive such causal relations, we can only use *observational* data collected from a patient population of interest. Motivated by the fact that the majority of patients only have a subset of genes sequenced[1], in Chapter 3, we consider the benefit of incorporating a large *confounded* observational dataset (*confounder unobserved*) alongside a small *deconfounded* observational dataset (*confounder revealed*) when estimating the *average treatment effect* (ATE). Our theoretical results show that the inclusion of confounded data can significantly reduce the quantity of deconfounded data required to estimate the ATE to within a desired accuracy level. Moreover, in some cases—say, genetics—we could imagine retrospectively selecting samples to deconfound. We demonstrate that by actively selecting these samples based upon the (already observed) treatment and outcome, we can reduce our data dependence further. Our theoretical results establish that the worst-case relative performance of our approach (vs. random selection) is bounded while our best-case gains are unbounded. We perform extensive synthetic experiments to validate our theoretical results. Finally, we demonstrate the practical benefits of selective deconfounding using a large real-world dataset related to genetic mutation in cancer.

Chapter 4 focuses on liquid biopsies—simple blood tests that can be used for accurate early stage cancer detection. In particular, we study a set of problems that occur in the development of liquid biopsies via the lens of *active sequential hypothesis testing* (ASHT). In the problem of ASHT, a learner seeks to identify the *true* hypothesis from among a known set of hypotheses. The learner is given a set of actions and knows the random distribution of the outcome of any action under any true hypothesis. Given a target error $\delta > 0$, the goal is to sequentially select the fewest number of actions so as to identify the true hypothesis with probability at least $1 - \delta$. Motivated by applications in which the number of hypotheses or actions is massive (e.g., genomics-based cancer detection), we propose efficient (greedy, in fact) algorithms and provide the first approximation guarantees for ASHT, under two types of adaptivity. Both of our guarantees are independent of the number of actions and logarithmic in the number of hypotheses. We numerically evaluate the performance of our algorithms using both synthetic and real-world DNA mutation data, demonstrating that our algorithms outperform previously proposed heuristic policies by large

[1] Often this subset is the same across the patient population because doctors will only order a gene to be sequenced if there are known treatments for that gene.

margins.

Finally, Chapter 5 is an empirical chapter, where we focus on solving real-world problems where we collaborate with physicians. This chapter is motivated by improving the gap between machine learning research in healthcare and what has been implemented in practice. In particular, we collaborated closely with Dr. Patel and Dr. Novelli from University of Pittsburgh Medical Center in predicting the 30-day readmission risk for patients with sickle cell disease. Reducing preventable hospital readmissions in SCD could potentially improve outcomes and decrease healthcare costs. In a retrospective study of electronic health records, we hypothesized *machine learning* (ML) algorithms may outperform standard readmission scoring systems (LACE and HOSPITAL indices). Participants (n=446) included patients with SCD with at least one unplanned inpatient encounter between January 1, 2013, and November 1, 2018. Patients were randomly partitioned into training and testing groups. Unplanned hospital admissions (n=3299) were stratified to training and testing samples. Potential predictors (n=486), measured from the last unplanned inpatient discharge to the current unplanned inpatient visit, were obtained via both data-driven methods and clinical knowledge. Three standard ML algorithms, *logistic regression* (LR), *support vector machine* (SVM), and *random forest* (RF) were applied. Prediction performance was assessed using the C-statistic, sensitivity, and specificity. In this dataset, ML algorithms outperformed LACE (C-statistic 0.6, 95%CI 0.57-0.64) and HOSPITAL (C-statistic 0.69, 95%CI 0.66-0.72), with the RF (C-statistic 0.77, 95%CI 0.73-0.79) and LR (C-statistic 0.77, 95%CI 0.73-0.8) performing the best. We reported the most important predictors in our best models, and derive clinical insights.

1.1 Review of Classical Results in Concentration Inequalities for Subgaussian Random Variables

We first state some classic results that we will use frequently in Chapters 3 and 4. We will consider the commonly used subgaussian distributions (Vershynin 2018). Loosely speaking, a random variable is subgaussian if its tail vanishes at a rate faster than some Gaussian distributions.

Definition 1 (subgaussian norm). *Let X be a random variable, its subgaussian norm is defined as $\|X\|_{\psi_2} := \inf\{t : \mathbb{E}[e^{X^2/t^2}] \leq 2\}$. Moreover, X is called subgaussian if $\|X\|_{\psi_2} < \infty$.*

Many commonly used distributions satisfy this assumption, e.g., Bernoulli, uniform, and Gaussian distributions etc. We introduce a standard concentration bound for subgaussian random variables.

Theorem 1 (Hoeffding Inequality Vershynin 2018). *Let $X_1, ..., X_n$ be independent subgaussian* *random variables. Then for any $\eta > 0$, it holds that*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}X_i\right| \geq \eta\right] \leq 2\exp\left(-\frac{2\eta^2}{\sum_{i=1}^{n} \|X_i\|_{\psi_2}^2}\right).$$

1.1.1 Special Case: Bounded Random Variables

In addition to being subgaussian, if the random variable is bounded, then a stronger version of Theorem 1 can be stated as follows:

Lemma 1 (Hoeffding's Lemma)**.** *Let X be any real-valued random variable with expected value $\mathbb{E}[X] = 0$, such that $a \leq X \leq b$ almost surely. Then, for all $\lambda \in R$, $\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$.*

Theorem 2 (Hoeffding's inequality for general bounded r.v.s)**.** *Let $X_1, ..., X_N$ be independent random variables such that $X_i \in [m_i, M_i], \forall i$. Then, for $t > 0$, we have $P\left(\left|\sum_{i=1}^{N} (X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right)$.*

Chapter 2

Personalized Treatment for Opioid Use Disorder

2.1 Introduction

The national opioid use disorder crisis in the United States leads to thousands of death annually (Skolnick 2018), affecting populations from all demographics. Since repeated opioid use can alter how we perceive motivation and reward long-term (Humphreys et al. 2017), making a full recovery from OUD is difficult and typically costly. In the United States, *medication assisted treatment* (MAT)—the standard treatment for OUD—includes the use of medications in combination with counseling and behavioral therapies. Effective treatment is made difficult by the fact that patients react to medications for OUD differently, for example, due to genetic variations (Crist et al. 2013), cultural or ethnic differences (Campbell and Edwards 2012), and stress (Sinha 2008). As a result, the treatment retention rate among patients with OUD remains low while the death rate remains high: studies have shown that the 2-month retention rate among patients with OUD is 57% with few staying enrolled beyond 3 months (Skolnick 2018), and the 30-year (after the initialization of the first treatments) death rate is 47% (Grella and Lovinger 2011). Thus, there is a need to develop better, personalized, treatment for OUD.

One key step in developing personalized treatment for OUD is to measure patients' treatment responses. While most outpatient programs estimate the effectiveness of a treatment through relapse rate (via urine tests), *ecological momentary assessment* (EMA) studies—in which patients respond to daily surveys—find that addressing craving episodes *before relapse* can likely help prevent actual relapse (Serre et al. (2012, 2015), Epstein et al. (2009)). However, EMA is not reliable in detecting cravings because it is subject to response bias: for example, adolescents tend to provide random information (McLellan et al. 1992), and patients ashamed about cravings may provide falsified information (Kleber et al. 2006).

In this work, we provide a framework to investigate the benefit of incorporating wearable devices in treatments for OUD, where wearable devices are defined as smart electronic devices that can be worn on wrists to collect data. There is an emerging trend of integrating wearable devices in medical treatments (Cheol Jeong et al. 2018), and pilot studies demonstrate that wearables can be useful in treating Parkinson's disease (Suzuki et al. 2017), post-discharge monitoring of ICU patients (Kroll et al. 2017), and detecting early-stage Alzheimer's disease (Varatharajan et al. 2019). Moreover, Fatseas et al. (2011, 2015) show that relapses and strong cravings could potentially be captured by the sensors contained inside the wearables. As a result, many start-ups (Valant et al. 2018, Linder 2019) are integrating wearables into treatments for OUD, and there is thus an urgent need to develop tools for assessing the value of wearable devices in those treatments. Throughout this work, we assume there exists an algorithm that detects patient health states in real-time, possibly with some uncertainty, using the features captured by wearables.

However, the potential advantage of wearables is constrained by the limited national budget to fight this epidemic (NIDA 2020). Specifically, it is unclear whether reducing the amount of money spent on MAT in favor of buying wearables is cost-effective. Complicating this problem, there is a variety of wearables with different prices, sensors, and accuracies available. In this work, we provide a framework to assess the cost-benefit trade-off of different wearables from the perspective of the healthcare system, to help determine whether and which wearables should be invested in by treatment programs for OUD.

Our contributions in this research are threefold. First, we provide a framework for understanding the values of different wearable devices in OUD treatments under budget constraints. Since patient health states are not always fully observable, we formulate a budget-constrained, discrete-time, finite-horizon, non-stationary partially observable Markov decision process. We consider three classes of MAT treatments in addition to counseling only and no treatment, incorporating different transition matrices to model cases in which we have no wearables, have wearables that provide different levels of information accuracy on patient health states, or have wearables that provide perfect information, i.e., a full information MDP benchmark model.

Second, we provide a novel budget reformulation for our CPOMDP that could potentially be applied to other CPOMDP models. To our knowledge, only two works have proposed methods to solve finite-horizon CPOMDPs (Cevik et al. 2018, Undurti and How 2010). However, the former has no feasibility guarantees, and neither have optimality guarantees. In contrast, our budget reformulation finds all optimal solutions lying on the convex hull of the original formulation's solutions. Moreover, our reformulation can be solved using a binary search in conjunction with an exact POMDP algorithm. (Similarly, in the case of CMDP, our formulation can be solved using a binary search in conjunction with an exact MDP algorithm.) We show that our reformulation not only guarantees the feasibility of our solution, but also optimality when randomized policies are allowed.

Third, we conduct a numerical study from the perspective of the healthcare system to provide insights that could potentially guide future field studies. Incorporating different device costs and accuracies, our objective is to maximize the total lifetime discounted *quality-adjusted life days* (QALDs) of patients subject to the budget constraint. We discover that the health benefit of incorporating wearables could be significant when the budget is not very generous, because if the budget is very generous, we can afford the most expensive (and effective) treatment in every period, and the marginal benefit of wearables is negligible. Furthermore, assuming that wearables do not affect patient treatment adherence levels, different patient types obtain different levels of benefit from wearables: patients with the highest treatment adherence benefit the least from wearable devices at all budget levels, and patients with lower TAs benefit the most when the budget is relatively low.

The paper is organized as follows. After a literature review in § 2.2, we formulate our model in § 2.4. In § 2.5 we introduce our budget reformulation, review an exact solution method for solving unconstrained POMDPs, and provide algorithms to solve our reformulation. In addition, we develop and introduce a heuristic algorithm to speed up the solving time of our reformulation, and provide a worst case error bound in § 2.15. In § 2.6 we compare wearables with various features and accuracies, and discuss our numerical results. We conclude in § 2.7. A notation table is included in § 2.3.

2.2 Literature Review

OUD Treatments In addition to the reasons listed in the introduction (genetics, cultural and ethnic differences, and stress), the effectiveness of a treatment for OUD can be influenced by comorbid medical conditions (Luo and Levin 2017), age (McLellan et al. 1994), co-occurring mental health disorders (Morse and Bride 2017), education levels, psychiatric functioning, marital separation or social functioning (Sayre et al. 2002), treatment enrollment duration (Eastwood et al. 2017), and multi-drug usage (Williamson et al. 2006). The breath of these features highlights the potential of personalized treatments.

Many tools have been investigated to try to reverse the opioid epidemic. These include strengthening the regulations for opioid prescription (Kolodny and Frieden 2017), predicting opioid overdose via machine learning (Lo-Ciganic et al. 2019), and increasing the accessibility of OUD treatments (Marshall et al. 2015). However, none address efficient treatments. Several studies established that some treatment for OUD is more cost-effective than no treatment, due to reduced hospital visits (e.g. Baser et al. (2011)), using either statistical tools or simulation models.

None explored the use of personalized treatment.

MDP and POMDP in Personalized Treatment A few works in the Operations Research/-Management Science literature have considered budget constraints in treatment decision models. Ayvaci et al. (2012) and Cevik et al. (2018) used finite horizon MDP and POMDP respectively to model the optimal breast cancer screening policy under budget constraints. Chen et al. (2018) formulated a finite horizon POMDP to model optimal liver cancer screening policies for patients with hepatitis C–infection under the constraint that the policies can change at most a given number of times, and conducted numerical experiments using a MDP. All three papers reformulated their problems into *mixed integer linear programs* (MILPs). Ayvaci et al. (2012) showed that the optimal patient health outcome is strictly concave with respect to budget if randomized policies were allowed; to enforce a deterministic optimal solution, they must add integrality constraints in their MILP model. Since the space of reachable belief states in a POMDP model is infinite, the reformulation proposed by Cevik et al. (2018) is computationally intractable. As a result, they discretized their belief space and obtained an approximate solution. Chen et al. (2018) showed that the marginal benefit of surveillance is higher in patient populations with faster disease progressions. Furthermore, as patients' risk of developing cancer diminishes and their health outcome improves, the frequency of surveillance should not increase.

Additional studies have used either an MDP or a POMDP to model clinical decisions, with the majority focusing on maximizing *quality-adjusted life years* (QALYs). Zhang et al. (2012) studied optimal prostate biopsy referral decisions using a infinite-horizon, non-stationary POMDP. Ayer et al. (2012, 2015) and Alagoz et al. (2013) studied the optimal clinical decisions related to breast cancer using finite-horizon POMDPs, and MDPs respectively. Erenay et al. (2014) and Suen et al. (2017) studied optimal colonoscopy screening and optimal drug sensitivity test in tuberculosis treatment, respectively, using finite-horizon, non-stationary POMDPs.

CPOMDPs in the Computer Science Literature Incorporating a cost constraint directly into a POMDP yields a model that is computational intractable. Two lines of work in the computer science literature address this problem in the setting of infinite horizon CPOMDPs and finite horizon CPOMDPs, respectively, with the majority focusing on the former. Both directions heavily rely on reformulating the problem into either an MILP or a *linear program* (LP). Within the infinite horizon setting, Isom et al. (2008) modified the pruning step in an exact algorithm for solving unconstrained POMDPs to incorporate the constraint. Kim et al. (2011) proposed a heuristic that uses *point-based value iteration* in conjunction with an LP. Poupart et al. (2015) converted the problem into an LP and considered only a subset of belief states. To ensure optimality, they proposed an iterative algorithm to enlarge the subset of the belief states. However, this method cannot be adopted into the finite-horizon setting. Lee et al. (2018) assumed unknown transition matrices and proposed to solve an unconstrained POMDP while optimizing its LP-induced parameters that control the trade-off between rewards and costs. None of the methods above have feasibility guarantees, and the majority do not guarantee that the final solution can be made sufficiently close to the optimal solution. In the finite horizon setting, Undurti and How (2010) proposed an algorithm that combines an offline finite lookahead and an online branch-and-bound algorithm to ensure the feasibility of the solution, however without optimality guarantees. Furthermore, to calculate the expected reward and cost, all methods mentioned above must either solve a system of linear programs or conduct simulations.

In contrast, our finite-horizon budget reformulation guarantees the feasibility of the final solution and admits optimal solutions when randomized policies are allowed; when deterministic solutions must be enforced, our reformulation might find a sub-optimal solution. In addition, we show that our reformulation can be solved using a binary search in conjunction with an exact POMDP (or MDP) algorithm.

2.3 Notation Table

Notation	Description
TA	treatment adherence
MAT	medication assisted treatment
EMA	ecological momentary assessment
QALD	quality-adjusted life days
QALY	quality-adjusted life years
M	*methadone* maintenance treatments with counseling
B	*buprenorphine* maintenance treatments with counseling
NT; IN; CO	no treatment; implant naltrexone with counseling; counseling only
Re, Dx, OD, Dt	states of replase, detoxification, overdose, and death respectively
NC, C1, C2	states of no craving, low craving, and high craving intensity, respectively
Abs	absorbing state
ATD	average treatment dynamics
PTD	personalized treatment dynamics
UT	urine test
N	number of treatment periods
S	the set of all information states
A	the set of all actions

O	the set of all observations
$\mathcal{P}(s_{t+1}\|s_t, a)$	transition probability under action a at time t
$\mathcal{P}_t^a$	transition probability matrix under action a at time t
$w(o_t\|s_t, a)$; W	observation probability; matrix containing $w(o_t\|s_t, a)$
σ; $\bar{\sigma}$	sensitivity ; 1− sensitivity
p; $\bar{p}$	specificity ; 1− specificity
β; B	belief state; the set of all belief states
$h_t^a(s)$, $c_t^a(s)$	health reward and cost at state s under action a at time t, respectively
$r_t(a, s)$	immediate reward at state s under action a at time t
Γ_t	amount of budget allocated for the rest of horizon in month t
$V_t^*(\beta_t)$	optimal expected value of the objective function at time t under belief β_t
$\mathcal{P}(o_{t+1}\|a, \beta_t)$	probability of observing o_{t+1} after taking action a at belief state β_t
Π	set of all feasible policies
$\pi'(\Gamma_t)$	optimal treatment policy under budget Γ_t (in System (1))
β_0	initial belief state
$H_{\beta_t}^{\pi}$,$C_{\beta_t}^{\pi}$	expected health and cost under belief β_t and policy π, respectively
θ	tunable parameter that takes values between 0 and 1
$V^{\pi}(\beta_t, \theta)$	expected reward under policy π, belief β_t, and θ
π^*	optimal policy in System (2)
θ^*	optimal θ in System (2)
Π_θ^*	optimal solution set to the unconstrained POMDP in System (2) under θ
π_θ^*	solution with the lowest expected health in Π_θ^*
$V^{\pi^*}(\beta_0, \Gamma_t)$	optimal value in System (2) under budget Γ_t
$\pi_{\theta^*}^*$	optimal policy with the lowest expected health under θ^* in Equation (2.9)
$\hat{F}(\Gamma_t)$	obtained by connecting the end points of the step function, $H_{\beta_0}^{\pi'}(\Gamma_t)$
$F(\Gamma_t)$	point-wise smallest concave function whose hypograph contains that of $\hat{F}(\Gamma_t)$
α_t	vector containing the expected reward at each state s_t
$\mathscr{A}_t^{a,o}$	minimal representation of the α-vector set for action a_t, observation o_{t+1}
$\mathscr{A}_t^a$	minimal representation of $\bigcup_{o_{t+1}\in O_{t+1}} \mathscr{A}_t^{a,o}$
$\mathscr{A}_t$	minimal representation of $\bigcup_{a_t\in A_t} \mathscr{A}_t^a$
$\tau(\alpha_{t+1}, a_t, o_{t+1})(s_t)$	scaled expected reward at state s_t given α_{t+1} after taking a_t and observing o_{t+1}
χ_t	the vector containing the expected cost at each state s_t
x_s, nc$_s$, c_s^2,	probability of transitioning to Dx, NC, C2 from state s, respectively
e_s, od$_s$, d_s	probability of transitioning to Re, OD, Dt from state s, respectively
w	probability of withdrawing
$\{T_0^a\}_{a\in A}$	transition matrices for ATD

$\{T_1^a\}_{a \in A}$	transition matrices for PTD

2.4 Constrained Partially Observable Markov Decision Process

In this section we describe our CPOMDP model, where a CPOMDP is defined as a POMDP with two additional components (Isom et al. 2008): 1) a cost incurred in each state for executing an action, and 2) an upper bound on the cumulative cost. We consider three different cases: (1) without any wearables; (2) with wearable devices that can detect cravings with various levels of accuracy; and (3) with wearable devices that can capture craving episodes with perfect accuracy. For each case, the objective of our model is to maximize the QALDs for an individual patient subject to a predefined budget constraint for the patient (which would be a fraction of the overall budget).

Time Horizon Let N denote the number of treatment periods. Although our model can be solved for longer horizons, to keep our model parsimonious, we use a one-year horizon to mimic the federal budget allocation for treating OUD. We further discretize the horizon into twelve months ($N = 12$)—the recommended change in treatments by the American Psychiatric Association (Kleber et al. 2006) is less than once per month. Within each month, a patient can transition between different states. Because the natural granularity of the clinical data makes daily treatment transition parameters easier to define and estimate than monthly ones, our transition probabilities and immediate rewards are defined in days.

Actions According to the American Society of Addiction Medicine (ASAM 2016), methadone, buprenorphine, and naltrexone are three standard medical treatments for OUD; these treatments are typically provided along with counseling and other support. Therefore, at the beginning of each month, a care provider can decide which one of the following five actions to take: *no treatment* (NT), *methadone maintenance treatment with counseling* (M), *buprenorphine maintenance treatment with counseling* (B), *implant naltrexone with counseling* (IN), or *counseling only* (CO). Because treatments M and B include prescribing medicine on a daily basis, their treatment outcomes are positively correlated with patient treatment adherence levels. Treatment IN requires only monthly implant procedures, and thus it works the same for all TA groups. Several treatment constraints mentioned by Kleber et al. (2006) are not implemented in our model but can be added easily: for example, treatment B is not suitable for patients with liver disease.

States The information states (S) in our models correspond to patient health states. We do not define the state recovery in our model because OUD is unlikely to be cured within one year: from our conversion with practitioners, a patient faces the risk of relapse even after 10 years of abstinence. We assume that if a patient starts to use drugs again—*relapses* (Re)—in the program, we will allow him or her to stay in the program. If a patient relapses, he or she can either go through a *detoxification* (Dx) program to stop using drugs or stop on his or her own (Zarkin et al. 2005). Because our optimization problem terminates once a patient withdraws from the program, we create an *absorbing state* (Abs) representing that the patient has either withdrawn from the program or died. In any given in-treatment day, (i.e., a day outside of states Re, Dx, or Abs,) a patient can experience either *no craving* (NC) for drugs the entire day or some craving at certain points of the day; clinical papers have found that not all cravings lead to relapse (Marsden et al. 2014, Serre et al. 2018). Based on conversations with practitioners, we define a *low craving intensity state* (C1) and a *high craving intensity state* (C2); the patient is more likely to relapse in state C2 than in C1, and in state C1 than in NC. Any renewed opioid usage after a period of abstinence carries an increased risk of *overdose* (OD) requiring medical attention due to loss of tolerance (Chalana et al. 2016), which can lead to *death* (Dt). Thus, the health of a patient falls into one of the following eight states, $\{\text{Dx}, \text{NC}, \text{C1}, \text{C2}, \text{Re}, \text{OD}, \text{Dt}, \text{Abs}\} := S$.

Transition Probability Matrices The transition probability $\mathcal{P}(s_{t+1}|s_t, a)$ is the transition probability from state $s_t \in S$ to state $s_{t+1} \in S$ under action $a \in A := \{\text{NT}, \text{M}, \text{B}, \text{IN}, \text{CO}\}$, where t indicates the number of days since the last (known) drug-use.[1] To reflect different wearable accuracies in detecting cravings and estimating individual reactions to treatments, we perturb the observation and transition matrices governing our POMDP, respectively. We describe how the transition probabilities were estimated or generated in § 2.6.2.

Observations and Observation Matrices Let $o_t \in O$ denote the observation at time t. In our model, the set of feasible observations, $O = \{\text{Dx}, \text{NC}, \text{C1}, \text{C2}, \text{Re}, \text{OD}, \text{Dt}, \text{Abs}\}$[2], is the same for all actions at all t. At every period, we perform urine test to decide whether a patient has relapsed or not. Since this assumption is potentially more conservative than necessary (as we could reduce the frequency of urine tests if the wearable is sufficiently accurate), we relax this assumption in § 2.6.4. We assume that urine tests can accurately detect drug usage within a three-day interval (Lautieri 2019). Let $w(o_t|s_t, a)$ be the probability of making observation o_t in state s_t under action a, and

[1] To keep the our model Markovian, the transition probability out of the state detoxification was modeled as a geometric distribution as indicated by Table 2.6.

[2] Note that as we will see in § 2.6, it is important that $|O|$ is the same as $|S|$ for the consistency of the model evaluation.

let W denote the matrix containing $w(o_t|s_t, a)$, with columns corresponding to observations and rows to true states.

Case 1 no wearables: When there is no wearable, we only observe either a *negative urine test result*, ut−, or a *positive result*, ut+. Since we do not observe patients' craving states—states NC, C1, and C2—we maintain a uniform belief over these three states if we observe a negative urine test. However, a positive urine test result does not necessarily indicate that the patient is in the state relapse because a patient can stop using drugs on his or her own. Thus, a care provider can partially observe state Re and has no information about state NC, C1, and C2 in this model. Let $\mathrm{nc}_{\mathrm{ut+}}$, $\mathrm{C1}_{\mathrm{ut+}}$, and $\mathrm{C2}_{\mathrm{ut+}}$ denote the probability of observing a positive urine test when the patient is in fact in states NC, C1, and C2, respectively. Then,

$$W = \begin{array}{c} \\ \text{Dx} \\ \text{NC} \\ \text{C1} \\ \text{C2} \\ \text{Re} \\ \text{OD} \\ \text{Dt} \\ \text{Abs} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{Dx} & \text{NC} & \text{C1} & \text{C2} & \text{Re} & \text{OD} & \text{Dt} & \text{Abs} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & (1-\mathrm{nc}_{\mathrm{ut+}})/3 & (1-\mathrm{nc}_{\mathrm{ut+}})/3 & (1-\mathrm{nc}_{\mathrm{ut+}})/3 & \mathrm{nc}_{\mathrm{ut+}} & 0 & 0 & 0 \\ 0 & (1-\mathrm{C1}_{\mathrm{ut+}})/3 & (1-\mathrm{C1}_{\mathrm{ut+}})/3 & (1-\mathrm{C1}_{\mathrm{ut+}})/3 & \mathrm{C1}_{\mathrm{ut+}} & 0 & 0 & 0 \\ 0 & (1-\mathrm{C2}_{\mathrm{ut+}})/3 & (1-\mathrm{C2}_{\mathrm{ut+}})/3 & (1-\mathrm{C2}_{\mathrm{ut+}})/3 & \mathrm{C2}_{\mathrm{ut+}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

The calculations of $\mathrm{nc}_{\mathrm{ut+}}$ and $\mathrm{C1}_{\mathrm{ut+}}$ are included in § 2.8. After plugging in our estimated transition matrices derived from past literature, $\mathrm{nc}_{\mathrm{ut+}}, \mathrm{C1}_{\mathrm{ut+}}, \mathrm{C2}_{\mathrm{ut+}} < 0.01$ since the probability that a patient recovers from a relapse within 3 days is very small (Zarkin et al. 2005). Therefore, the state Re is *fully observable* in this case.

Case 2 wearables with imperfect information: In this case, we assume there exists an algorithm that takes urine test results and data collected via the wearables as inputs and returns an estimate of the patient's current health state. However, depending on the accuracy and dimension of the inputs, the algorithm will have different sensitivities and specificities for each partially observable state, where the *sensitivity* (σ_s) is the probability of observing state $s \in S$ given that the patient is in state s, and the *specificity* (p_s) is the probability of not observing state s given that the patient is not in state s. To simplify the representation of W, we parametrize[3] W as follows, where we put more weights on worse health states:

[3] There are many equivalent parameterization of this problem.

$$W = \begin{array}{c|cccccccc} s/o & \text{Dx} & \text{NC} & \text{C1} & \text{C2} & \text{Re} & \text{OD} & \text{Dt} & \text{Abs} \\ \hline \text{Dx} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{NC} & 0 & p_{\text{Re}}p_{\text{C2}}p_{\text{C1}} & p_{\text{Re}}p_{\text{C2}}\bar{p}_{\text{C1}} & p_{\text{Re}}\bar{p}_{\text{C2}} & \bar{p}_{\text{Re}} & 0 & 0 & 0 \\ \text{C1} & 0 & p_{\text{Re}}p_{\text{C2}}\bar{\sigma}_{\text{C1}} & p_{\text{Re}}p_{\text{C2}}\sigma_{\text{C1}} & p_{\text{Re}}\bar{p}_{\text{C2}} & \bar{p}_{\text{Re}} & 0 & 0 & 0 \\ \text{C2} & 0 & p_{\text{Re}}\bar{\sigma}_{\text{C2}}p_{\text{C1}} & p_{\text{Re}}\bar{\sigma}_{\text{C2}}\bar{p}_{\text{C1}} & p_{\text{Re}}\sigma_{\text{C2}} & \bar{p}_{\text{Re}} & 0 & 0 & 0 \\ \text{Re} & 0 & \bar{\sigma}_{\text{Re}}p_{\text{C2}}p_{\text{C1}} & \bar{\sigma}_{\text{Re}}p_{\text{C2}}\bar{p}_{\text{C1}} & \bar{\sigma}_{\text{Re}}\bar{p}_{\text{C2}} & \sigma_{\text{Re}} & 0 & 0 & 0 \\ \text{OD} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \text{Dt} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \text{Abs} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array},$$

where $\bar{\sigma}_s = 1 - \sigma_s$, and $\bar{p}_s = 1 - p_s$, for $s \in \{\text{Re}, \text{C1}, \text{C2}\}$. As in the previous case, Re is fully observed if monthly urine tests are present: we set $p_{Re} = \sigma_{Re} = 1$.

Case 3 wearables with perfect information: In this case, wearables can correctly detect a patient's health state. Thus, the observation matrix is the identity matrix, and the model becomes a Markov decision process.

Belief States The belief state at time t, $\beta_t = (\beta_t(\text{Dx}), \beta_t(\text{NC}), \beta_t(\text{C1}), \beta_t(\text{C2}), \beta_t(\text{Re}), \beta_t(\text{OD}), \beta_t(\text{Dt}), \beta_t(\text{Abs})) \in B$, defines the probabilities that the care provider believes the patient is in before any action is taken. If the patient is in a fully observable state, s, then $\beta_t(s) = 1$ and $\beta_t(s') = 0$ for all $s' \neq s$. If a patient is in one of the partially observable states, i.e., states NC, C1, C2, and Re, the belief vector can be represented as $\beta_t = (0, \beta_t(\text{NC}), \beta_t(\text{C1}), \beta_t(\text{C2}), \beta_t(\text{Re}), 0, 0, 0)$, where $\beta_t(\text{NC}) + \beta_t(\text{C1}) + \beta_t(\text{C2}) + \beta_t(\text{Re}) = 1$.

Immediate Rewards In our CPOMDP, the immediate reward is the QALD. In this paper, we will use the terms *treatment outcome*, *QALDs*, and *health gain* interchangeably. To represent the QALDs, we assign a value, $h_t^a(s) \in [0, 1]$, in every period to every action and health state pair (a, s). That is, the immediate reward, $r_t(a, s)$, that a patient gains at state s under treatment decision a at time t is $h_t^a(s)$, for $t \in \{0, ..., N-1\}$[4]. We denote the terminal reward $r_N(a, s)$, which equals to $h_N(s)$ in our model. Therefore, the belief state immediate health reward under treatment a at time t is $\sum_{s \in S} r_t(a, s)\beta_t(s)$, and we denote this value $h_t^a(\beta_t)$. To avoid an overly myopic optimal policy, we calculate the expected health gain for the patient under no treatment from day 360 to day 420 and add it as the terminal reward. We discuss the significance and sensitivity of the terminal reward in § 2.6.4.

[4] In our model the immediate health reward, $h_t^a(s)$, is independent of t. That is $\forall t, h_t^a(s) = h^a(s)$. We include the parameter t here for generality.

Costs The expected cost at time t under treatment a at state s , $c_t^a(s)$, is always negative.[5] When $t \in \{0, ..., N-1\}$, $c_t^a(s)$ includes the cost of detoxification if the patient is in the state Dx, the cost of hospital visits if a patient overdoses (see § 2.9), and the cost of treatment if the patient is not in the states Dx and Abs. Note that there is no terminal expected cost under this formulation, i.e. $c_N^a(s) = 0$ for all states $s \in S$ and actions $a \in A$, since the cost that will incur next year should be separated from the cost that will incur this year. Thus, the belief state expected cost under treatment a at time $t \in \{0, ..., N-1\}$ is $\sum_{s \in S} c_t^a(s)\beta_t(s)$, and we denote this quantity $c_t^a(\beta_t)$.

Budget Constraint The budget constraint is incorporated via an open-loop optimization formulation. Let $\Gamma_t > 0$ denote the amount of budget that is allocated for the patient for the rest of the horizon, in month t. In each month t, we solve a new optimize problem using the budget Γ_t. To reflect the costs of wearable devices, we deduct the cost of the specific wearable from the annual budget before solving the optimization problem.

Optimality Equations The optimal treatment action sequence maximizes the expected reward gained throughout the planning horizon. The optimal solution can be solved using dynamic programming techniques (for example, see Cassandra et al. (1997)). Let $V_t^*(\beta_t)$ denote the optimal expected value of the objective function at time t under belief β_t. Let $\mathcal{P}(o_{t+1}|a, \beta_t)$ be the probability of making observation o_{t+1} after taking action a at belief state β_t, that is $\mathcal{P}(o_{t+1}|a, \beta_t) = \sum_{s_{t+1} \in S} w(o_{t+1}|s_{t+1}, a) \sum_{s_t \in S} \beta_t(s_t)\mathcal{P}(s_{t+1}|s_t, a)$. Let β_{t+1} be the updated belief given the old belief β_t, observation o_{t+1}, and action a. Then, the optimal Bellman's equation for an unconstrained POMDP satisfies:

$$V_t^*(\beta_t) = \max_{a \in A} \left\{ \sum_{s_t \in S} r_t(s_t, a)\beta_t(s_t) + \sum_{o_{t+1} \in O} \mathcal{P}(o_{t+1}|a, \beta_t) V_{t+1}^*(\beta_{t+1}) \right\} \tag{2.1}$$

$$= \max_{a \in A} \quad \sum_{s_t \in S} \beta_t(s_t) \left\{ r_t(s_t, a) + \sum_{o_{t+1} \in O} \sum_{s_{t+1} \in S} w(o_{t+1}|s_{t+1}, a)\mathcal{P}(s_{t+1}|s_t, a) V_{t+1}^*(\beta_{t+1}) \right\}. \tag{2.2}$$

We express β_{t+1} in terms of known model parameters, that is, the transition probability $\mathcal{P}(s_{t+1}|s_t, a)$ and observation probability $w(o_{t+1}|s_{t+1}, a)$:

$$\begin{aligned}
\beta_{t+1}(s_{t+1}) \ := \mathcal{P}(S_{t+1} = s|o_{t+1}, a, \beta_t) &= \frac{\mathcal{P}(o_{t+1}, s|a, \beta_t)}{\mathcal{P}(o_{t+1}|a, \beta_t)} = \frac{\mathcal{P}(o_{t+1}|s, a, \beta_t)\mathcal{P}(s|a, \beta_t)}{\mathcal{P}(o_{t+1}|a, \beta_t)} \\
&= \frac{w(o_{t+1}|s, a) \sum_{s' \in S} \mathcal{P}(S_{t+1} = s, S_t = s'|a, \beta_t)}{\mathcal{P}(o_{t+1}|a, \beta_t)} \\
&= \frac{w(o_{t+1}|s, a) \sum_{s' \in S} \mathcal{P}(s|s', a)\beta_t(s')}{\mathcal{P}(o_{t+1}|a, \beta_t)}. \qquad (2.3)
\end{aligned}$$

[5] Similarly, the expected cost, $c_t^a(s)$, in our model is also independent of t.

The second to the last equality is because the observation o_t is independent of the belief β_t. Note that the belief vector in conjunction with the belief update absorb the entire history of the model and thus achieve the Markovian property (Smallwood and Sondik 1973).

Constrained POMDP Model The goal of the constrained POMDP model is to find the optimal treatment plan, $\pi'(\Gamma_t) \in \Pi$, that yields the maximum expected reward through the planning horizon while satisfying the budget constraint, Γ_t, where Π denotes the set of of all feasible policies. Recall $h_t^a(\beta_t)$ is the expected health reward of taking treatment a at time t under the clinicians' belief about patient's health state β_t, and let $H_{\beta_0}^{\pi}$ denote the expected health under an initial belief β_0 and policy $\pi \in \Pi$ in our CPOMDP problem, i.e.,

$$H_{\beta_0}^{\pi} = \mathbb{E}_{\beta_0}^{\pi}\left[\sum_{t=0}^{N-1} h_t^a(\beta_t) + h_N(\beta_t)\right]. \tag{2.4}$$

Unless otherwise mentioned, we fix this initial belief β_0 throughout the rest of the paper, and note that $H_{\beta_0}^{\pi}$ is always non-negative. Similarly, let $C_{\beta_0}^{\pi}$ be the expected cost under an initial belief β_0 and the policy π in our CPOMDP problem, i.e.

$$C_{\beta_0}^{\pi} = \mathbb{E}_{\beta_0}^{\pi}\left[\sum_{t=0}^{N-1} c_t^a(\beta_t)\right]. \tag{2.5}$$

Note that $C_{\beta_0}^{\pi}$ is always non-positive. We formulate the following optimization problem:

$$\textbf{System I: } \pi'(\Gamma_t) = \arg\max_{\pi\in\Pi} H_{\beta_0}^{\pi} \tag{2.6}$$

$$s.t. - C_{\beta_0}^{\pi} \leq \Gamma_t. \tag{2.7}$$

Constraint (2.7) guarantees that the expected cost of the treatment will be less than or equal to the budget. Since we model the problem as an open loop problem (we reoptimize the problem using an updated budget at each time step), it is guaranteed that our final solution would satisfy the budget constraint if we could solve the problem using System I.

Let $\{\pi'(\Gamma_t)\}_{\Gamma_t\in(0,\infty)}$ be the set of optimal solutions in System I when we vary Γ_t from 0 to ∞. Let $H_{\beta_0}^{\pi'}(\Gamma_t)$ denote the expected health under policy $\pi'(\Gamma_t)$. Before discussing the tractability of System I, we first list some properties that this optimal solution set satisfies:

Proposition 1. *Properties of the optimal solution set in System I:*

1(a) *The optimal policy $\pi'(\Gamma_t)$ is not necessarily unique, but the optimal expected health $H_{\beta_0}^{\pi'}(\Gamma_t)$ is unique for any fixed Γ_t and β_0.*

1(b) *The optimal expected health $H_{\beta_0}^{\pi'}(\Gamma_t)$ is non-decreasing in Γ_t.*

1(c) *Let $\mathcal{H}^{\pi'}_{\beta_0}$ denote the unique elements contained in the optimal solution set $\left\{ H^{\pi'}_{\beta_0}(\Gamma_t) \right\}_{\Gamma_t \in (0,\infty)}$. Then, the set $\mathcal{H}^{\pi'}_{\beta_0}$ is finite.*

Property 1(a) follows directly from uniqueness of Equation (2.6). Property 1(b) holds because the current optimal solution remains feasible after budget increase. Property 1(c) holds because the number of actions we can perform is finite and the initial belief β_0 is fixed.

The objective function, Equation (2.6), can be solved through exact POMDP algorithms using Equations (2.1)–(2.2). However, with Constraint (2.7), System I is numerically intractable to solve (Poupart et al. 2015). To address this problem, we provide a novel reformulation of our CPOMDP problem by incorporating the budget constraint into the objective function in the next section, and show that our reformulation can be solved using exact POMDP algorithms in conjunction with a binary search. When randomized policies are allowed, our reformulation solves the original problem exactly, but when deterministic policies are enforced, our reformulation might find a suboptimal solution.

2.5 Analytical Results

In this section we present our reformulation of the CPOMDP problem. The key idea in our budget reformulation is to integrate the budget constraint into the objective function through a tunable parameter, and optimize over this parameter. Mathematically, the immediate reward becomes a convex combination of the expected health and cost in our reformulation. Thus, our immediate reward at time t under treatment a at state s, $r_t(a, s)$, becomes $\theta h^a_t(s) + (1-\theta)c^a_t(s)$, where $\theta \in [0, 1]$, $h^a_t(s) \in [0, 1]$ and $c^a_t(s) < 0$, for $t \in \{0, ..., N-1\}$, and the terminal reward, $r_N(a, s)$, becomes $\theta h_N(s)$. The belief state immediate reward under treatment a, $r_t(a, \beta_t)$, takes the following form:

$$r_t(a, \beta_t) = \sum_{s \in S} r_t(a, s)\beta_t(s) = \theta h^a_t(\beta_t) + (1-\theta)c^a_t(\beta_t),$$

for $t \in \{0, ..., N-1\}$, and the belief state terminal reward is $r_N(\beta_N) = \theta h_N(\beta_N)$. Let $V^\pi(\beta_0, \theta)$ denote the expected reward function under a policy π, initial belief β_0, and parameter θ. By the linearity of expectation, we have the following Lemma (whose proof is included in § 2.10):

Lemma 2. *If the immediate reward and terminal reward of a discrete time finite horizon POMDP takes the form $\theta h^a_t(\beta_t) + (1-\theta)c^a_t(\beta_t)$ for every time step, then $V^\pi(\beta_0, \theta)$ can be written as $V^\pi(\beta_0, \theta) = \theta H^\pi_{\beta_0} + (1-\theta)C^\pi_{\beta_0}$.*

The rest of this section is organized as follows: we first state our CPOMDP reformulation in § 2.5.1 and then show the correctness of our reformulation in § 2.5.2 by comparing the solutions obtained in System I and our reformulation. We will then review *incremental pruning*—one of the exact solutions for unconstrained POMDP—and show how to use it in conjunction with a binary search to solve our reformulation in § 2.5.3.

2.5.1 CPOMDP Reformulation

For a fixed initial belief β_0, our budget reformulation comprises the following optimization system:

$$\textbf{System II: } \pi^* = \arg\max_{\pi \in \Pi^*_{\theta^*}} \left\{ -C^{\pi}_{\beta_0} \ : \ -C^{\pi}_{\beta_0} \leq \Gamma_t \right\} \tag{2.8}$$

$$\theta^* = \max \left\{ \theta \ : \ -C^{\pi^*_\theta}_{\beta_0} \leq \Gamma_t, \theta \in [0,1] \right\} \tag{2.9}$$

$$\pi^*_\theta = \arg\min_{\pi \in \Pi^*_\theta} H^{\pi}_{\beta_0} \tag{2.10}$$

$$\Pi^*_\theta = \arg\max_{\pi \in \Pi} \left\{ \theta H^{\pi}_{\beta_0} + (1-\theta) C^{\pi}_{\beta_0} \right\}, \tag{2.11}$$

and we denote the optimal policy under budget Γ_t in System II by $\pi^*(\Gamma_t)$. System II describes a two-step optimization problem. First, in Equation (2.11), for a fixed θ, we find the set of optimal policies, Π^*_θ, that finds the maximum expected reward throughout the planning horizon. In Equation (2.10), we pick the π^*_θ that has the smallest expected health $H^{\pi^*_\theta}_{\beta_0}$. Since the expected cost is negative, Equation (2.10) is also equivalent to finding the π^*_θ that yields the largest expected cost $C^{\pi^*_\theta}_{\beta_0}$. Second, in Equation (2.9), we find the largest θ, θ^*, such that the absolute expected cost that we find in Equation (2.11), $-C^{\pi^*_\theta}_{\beta_0}$, is under the budget, Γ_t. After finding θ^*, we resolve Equation (2.8) and denote this optimal policy π^*, inside $\Pi^*_{\theta^*}$ with the largest absolute cost such that the cost is under the budget constraint. Thus, we obtain the optimal solution $\pi^*(\Gamma_t)$. The purpose of Equations (2.10) and (2.8) is to handle the situations where Π^*_θ and $\Pi^*_{\theta^*}$ contain multiple optimal solutions. Since all policies inside $\Pi^*_{\theta^*}$ have the same objective function value, $V^\pi(\beta_0, \theta^*)$, maximizing the expected cost is equivalent to maximizing the expected health in Equation (2.8). In other words, Equation (2.8) is equivalent to $\arg\max_{\pi \in \Pi^*_{\theta^*}} \left\{ H^{\pi}_{\beta_0} \ : \ -C^{\pi}_{\beta_0} \leq \Gamma_t \right\}$.

Let $\{\pi^*(\Gamma_t)\}_{\Gamma_t \in (0,\infty)}$ denote the set of optimal policies in System II when we vary Γ_t from 0 to ∞, and let $H^{\pi^*}_{\beta_0}(\Gamma_t)$ denote the maximum expected health under policy $\pi^*(\Gamma_t)$. Let $V^{\pi^*}(\beta_0, \Gamma_t)$ denote the optimal value of System II under budget Γ_t. Let $\pi^*_{\theta^*}$ be the optimal policy with the lowest expected health under the optimal parameter θ^* as defined in Equation (2.9), i.e., $\pi^*_{\theta^*} = \arg\min_{\pi \in \Pi^*_{\theta^*}} H^{\pi}_{\beta_0}$. Then, similar to Proposition 1, we first list the properties that the optimal solutions in System II satisfy:

Proposition 2. *Properties of the solution set in System II:*

2*(a) The optimal policy $\pi^*(\Gamma_t)$ is not necessarily unique, but the values $H^{\pi^*}_{\beta_0}(\Gamma_t)$, $C^{\pi^*}_{\beta_0}(\Gamma_t)$ are unique for any fixed Γ_t.*

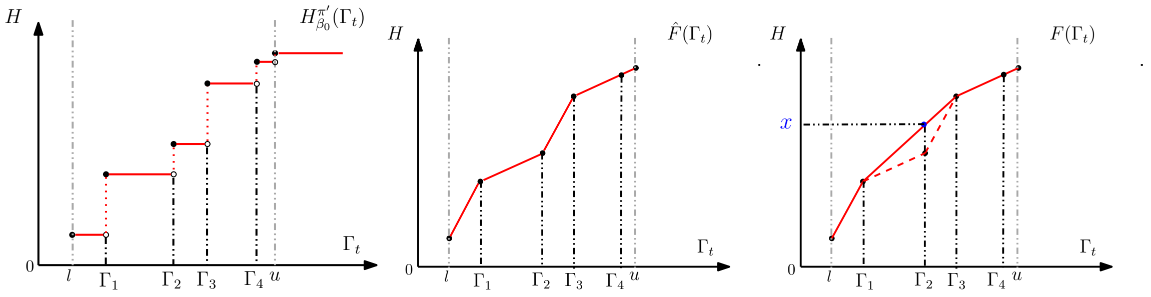


Figure 2.1: Construction of the efficiency frontier of the original formulation's feasible solution set convex hull

2*(b)* $H^{\pi^*_\theta}_{\beta_0}$, $-C^{\pi^*_\theta}_{\beta_0}$*, and the objective function* $\theta H^{\pi^*_\theta}_{\beta_0} + (1-\theta)C^{\pi^*_\theta}_{\beta_0}$ *are all non-decreasing in* θ*. Furthermore, as* Γ_t *increases,* θ^* *is non-decreasing.*

2*(c)* $H^{\pi^*}_{\beta_0}(\Gamma_t)$ *is non-decreasing in* Γ_t*.*

2*(d) If* $\gamma_2 > \gamma_1$ *and* $V^{\pi^*}(\beta_0, \gamma_2) > V^{\pi^*}(\beta_0, \gamma_1)$, $H^{\pi^*_{\theta^*}}_{\beta_0}(\gamma_2) > H^{\pi^*}_{\beta_0}(\gamma_1)$.

2*(e) Let* $\mathcal{H}^{\pi^*}_{\beta_0}$ *denote the unique elements contained in the set* $\left\{H^{\pi^*}_{\beta_0}(\Gamma_t)\right\}_{\Gamma_t \in (0,\infty)}$*. Then, the set* $\mathcal{H}^{\pi^*}_{\beta_0}$ *is finite.*

Proposition 2 shows that the optimal solution set in System II satisfies the properties of the optimal solutions in System I. Property 2(a) is directly implied by Equation (2.8). We prove Property 2(b) via construction, and the proof is included in § 2.11.1. Property 2(c) follows from Property 2(b), and the proof can be found in § 2.11.2. Property 2(d) states that under the same initial belief β_0, for any budget γ_2 that is sufficiently larger than budget γ_1 so as to obtain a strictly higher objective value, the lowest possible expected health among the set of optimal policies, $H^{\pi^*_{\theta^*}}_{\beta_0}(\gamma_2)$, is always greater than the highest expected health that we can obtain under budget γ_1, $H^{\pi^*}_{\beta_0}(\gamma_1)$. Property 2(d) is implied directly by Property 2(c) and also the proof of Property 2(b). After showing the correctness of our reformulation, Property 2(e) follows from Property 1(c) since the solutions that we find in System II is a subset of that in System I.

Note that our budget reformulation can be applied to any CPOMDP problem with one constraint, e.g., the *optimal breast cancer diagnose problem with budget constraint* proposed in Ayvaci et al. (2012), or the *quickest change detection* problem proposed in Isom et al. (2008). In addition, in § 2.13 we show that our reformulation can be extended to the case where we have multiple constraints. However, the running time of the current solution algorithm might grow exponentially as the number of constraints increases. Novel algorithms to solve the extended reformulation could be proposed, however this is beyond the scope of this paper.

2.5.2 Correctness of Our Reformulation

Proposition 1 suggests that, in System I, $H_{\beta_0}^{\pi'}(\Gamma_t)$ is a non-decreasing step function in Γ_t. Figure 2.1 (left) shows a plausible $H_{\beta_0}^{\pi'}(\Gamma_t)$ function, where each solid dot represents a set of optimal solutions that yield the same maximum expected health and the same expected cost which is under the budget Γ_t, and each solid line represents that the optimal solution set stays the same over an interval of budget values. Note that in System I it is possible that there are multiple solid dots on each red line segment, i.e., there exists multiple optimal solutions with different expected costs but the same expected health. However, this is not allowed in our reformulation, System II. This is because when there are multiple sets of solution with the same expected health but different expected costs, our objective function, Equation (2.11) will always prefer the solution set with the lowest expected cost.

If we connect the end points of this step function, $H_{\beta_0}^{\pi'}(\Gamma_t)$, then we obtain a piecewise linear, non-decreasing function $\hat{F}(\Gamma_t)$. Fig. 2.1 (middle) illustrates the construction of $\hat{F}(\Gamma_t)$ defined on $[l, u]$, where each point on the solid line represents a randomize policy that is a convex combination of the two nearest optimal solution sets (i.e., the two nearest solid dots). However, depending on the structure of each individual problem, this function $\hat{F}(\Gamma_t)$ is not necessarily concave (as in our illustrating example, the middle figure in Fig. 2.1). Let $F(\Gamma_t)$ be the point-wise smallest concave function whose hypograph contains the hypograph of $\hat{F}(\Gamma_t)$. Then, $F(\Gamma_t)$ is the efficiency frontier of the convex hull of the original formulation's (System I's) feasible solution set. That is, when randomized policies are allowed, the solutions that lie on $F(\Gamma_t)$ are optimal. If $\hat{F}(\Gamma_t)$ is concave, then $F(\Gamma_t) = \hat{F}(\Gamma_t)$, and if it is not, then when randomized policies are allowed, we will be able to find solutions lying on $F(\Gamma_t)$ that either dominate or are equal to the solutions lie on $\hat{F}(\Gamma_t)$. Fig. 2.1 (right) illustrates the construction of $F(\Gamma_t)$, where $F(\Gamma_t)$ is denoted with the solid curves. Indeed, this function $F(\Gamma_t)$ is piecewise linear, concave, and strictly increasing.

Next, in Theorem 3 we show that the our formulation, System II, recovers the convex hull of the feasible solution set in System I :

Theorem 3 (Correctness of our reformulation)**.** *Let $F(\Gamma_t)$ be piecewise linear, concave, and strictly increasing on $[l, u]$[6] as defined above, then $H_{\beta_0}^{\pi'} = H_{\beta_0}^{\pi^*}$ when $-C_{\beta_0}^{\pi'}$ lies on $F(\Gamma_t)$. That is, our reformulation, System II finds all solutions that lie on the efficiency frontier of the convex hull of the solutions of System I.*

The proof of Theorem 3 can be found in § 2.12. Theorem 3 implies that System II recovers the convex hull of the feasible solution set in our original formulation. That is, when randomized policies are allowed, our reformulation finds the exact solution to the original problem. However,

[6] We pick l, u such that F is $-\infty$ on $[0, l)$ (i.e., infeasible) and constant on $[u, \infty)$

when deterministic solutions are enforced, our reformulation will miss any optimal solutions that lie strictly within the convex hull of solutions. In our illustrating example, when deterministic policies are enforced, our System II will find the optimal policies that correspond to $l, \Gamma_1, \Gamma_3, \Gamma_4$, and u, but it will fail to find the optimal policy that corresponds to Γ_2 in Fig 2.1 (right). Note that Theorem 3 agrees with the findings in Ayvaci et al. (2012) where they show that when randomized policies are allowed, the optimal health outcome is strictly concave with respect to the budget. Although our reformulation provides stronger theoretical guarantees when randomized policies are allowed, to align our results with past literature on clinical decisions, we consider only deterministic optimal policies in our experiments.

2.5.3 Solution Method for Our Reformulation

In System II, we can compute Equations (2.8), (2.10), and (2.11) using the exact POMDP algorithm—*incremental pruning* (Zhang and Liu 1996, Cassandra et al. 1997). To solve for Equation (2.9), we observe that the monotonicity of $H_{\beta_0}^{\pi_\theta^*}$ and $-C_{\beta_0}^{\pi_\theta^*}$ with respect to θ in Property 2(b) implies that the feasible region of $\theta \in [0, 1]$ in Equation (2.9) is continuous, and thus we can solve for θ^* using a binary search in conjunction with any exact solution method for finite-horizon POMDPs. Before we provide a complete algorithm that solves our reformulation, we first review one of the exact finite-horizon POMDP solution methods that we will use—*incremental pruning*.

Review of Incremental Pruning—An Exact POMDP Solution Method The *incremental pruning* algorithm (Zhang and Liu 1996, Cassandra et al. 1997) can be applied to solve finite-horizon POMDPs exactly using backward induction; it relies heavily on the following decomposition of the optimal Bellman's equation (Equation (2.2)):

$$V_t^{a_t,o_{t+1}}(\beta_t) = \sum_{s_t \in S} \beta_t(s_t) \left\{ \frac{r(a_t, s_t)}{|O|} + \sum_{s_{t+1} \in S} w(o_{t+1}|s_{t+1}, a_t)\mathcal{P}(s_{t+1}|s_t, a)V_{t+1}^*(\beta_{t+1}) \right\}, \tag{2.12}$$

$$V_t^{a_t}(\beta_t) = \sum_{o_{t+1} \in O} V_t^{a_t}(o_{t+1}, \beta_t), \tag{2.13}$$

$$V_t^*(\beta_t) = \max_{a_t \in A} V_t^{a_t}(\beta_t). \tag{2.14}$$

In Equation (2.12), for a fixed belief β_t, we calculate the expected reward for each fixed observation o_{t+1} after we have taken action a_t, and we denote this value $V_t^{a_t,o_{t+1}}(\beta_t)$. In Equation (2.13), we then sum over all possible observations and thus obtain $V_t^{a_t}(\beta_t)$. Lastly, in Equation (2.14), we take the maximum over the set of all possible actions, selecting the action that yields the highest expected reward. One can easily verify that the above decomposition is equivalent to Equation (2.2).

To distinguish the belief state expected values from the expected value at each underlying state $s \in S$ at time t, we introduce *α-vector*s, α_t, to represent the vector containing the expected reward at each state s_t. Let $V_N(s_i)$ be the expected reward of state s_i at time N. Then,

$$\alpha_N \ := \langle V_N(s_1), ..., V_N(s_{|S|})\rangle. \tag{2.15}$$

Note that since we do not make decisions in period N, α_N is independent of actions. Thus, the belief state expected reward becomes $V_N(\beta_N) = \sum_{s_N \in S} \beta_N(s_N)V_N(s_N) = \beta_N \cdot \alpha_N$.

Since the transformations (2.12)–(2.14) preserve the piecewise linearity and convexity of the optimal Bellman's equation with respect to the belief β_t (Sondik 1971, Smallwood and Sondik 1973, Cassandra et al. 1997), there exists some unique finite parsimonious representations of the value functions $V_t^{a_t,o_{t+1}}$, $V_t^{a_t}$, and V_t^* (Sondik 1971, Zhang and Liu 1996). Let $\mathscr{A}_t^{a,o}$ denote the set of α*-vector*s under a fixed action a_t, observation o_{t+1} at time t. Let $\mathscr{A}_t^a$ be the set of α*-vector*s under a fixed action a_t at time t summed over the set of all possible observations $o_{t+1} \in O$ as described in Equation (2.13). Let $\mathscr{A}_t$ be the set of α*-vector*s at time t that includes all actions $a_t \in A$. We will provide the formal definition of those three variables below in Equations (2.20)–(2.22). Using induction, for some α*-vector* sets $\mathscr{A}_t^{a,o}$, $\mathscr{A}_t^a$, and $\mathscr{A}_t$, we can write

$$V_t^{a_t,o_{t+1}}(\beta_t) = \max_{\alpha \in \mathscr{A}_t^{a,o}} \beta_t \cdot \alpha, \tag{2.16}$$

$$V_t^{a_t}(\beta_t) = \max_{\alpha \in \mathscr{A}_t^{a}} \beta_t \cdot \alpha, \tag{2.17}$$

$$V_t^{*}(\beta_t) = \max_{\alpha \in \mathscr{A}_t} \beta_t \cdot \alpha. \tag{2.18}$$

Equations (2.16)–(2.18) can effectively reduce the number of vectors that we need to keep track of in Equations (2.12)–(2.14). So the idea of the algorithm is to find the set of α*-vector*s that are undominated at every belief state in each backward induction step.

Let $purge(\cdot)$ be the *minimal representation* of a set by removing the pointwise dominated vectors, and let $\oplus$ be the Minkowski sum of two sets[7]. Let $\alpha_{t+1}(s_{t+1})$ be the expected reward at state $t+1$ with $\alpha_{t+1} \in \mathscr{A}_{t+1}$, i.e., $\alpha_{t+1}(s_{t+1}) = V_{t+1}(s_{t+1})$. Then, we define $\tau(\alpha_{t+1}, a_t, o_{t+1})(s_t)$ to be the scaled expected reward at state s_t after taking action a_t and making the observation o_{t+1}, i.e.,

$$\tau(\alpha_{t+1}, a_t, o_{t+1})(s_t) = \frac{r(a_t, s_t)}{|O|} + \sum_{s_{t+1} \in S} w(o_{t+1}|s_{t+1}, a_t)\mathcal{P}(s_{t+1}|s_t, a_t)\alpha_{t+1}(s_{t+1}). \tag{2.19}$$

Thus, the minimal representation of the sets $\mathscr{A}_t^{a,o}$, $\mathscr{A}_t^a$, $\mathscr{A}_t$ at time t can be represented by

$$\mathscr{A}_t^{a,o} = purge\Big(\{\tau(\alpha, a, o)|\alpha \in \mathscr{A}_{t+1}\} \Big), \tag{2.20}$$

$$\mathscr{A}_t^{a} = purge\Bigg(\bigoplus_{o \in O_{t+1}} \mathscr{A}_t^{a,o} \Bigg), \tag{2.21}$$

$$\mathscr{A}_t = purge\Bigg(\bigcup_{a \in A_t} \mathscr{A}_t^{a} \Bigg). \tag{2.22}$$

[7] $X \oplus Y = \{x + y \ : \ x \in X, y \in Y\}$

Equation (2.20) corresponds to Equation (2.16), where we calculate the set of scaled expected reward vectors for all α-*vector*s obtained from the last iteration $\alpha \in \mathscr{A}_{t+1}$ using Equation (2.19), under a fixed observation o_{t+1} and action a. We then remove all vectors that are pointwise dominated by another vector inside this set as those vectors will never appear in the optimal solution. Similarly, Equations (2.21) and (2.22) correspond to Equations (2.17) and (2.18), where we sum over the set of all observations and combine the α-*vector*s under different actions respectively, and obtain the set of undominated vectors. Equation (2.21) can be solved efficiently using *incremental pruning* (Cassandra et al. 1997), and Equations (2.20) and (2.22) can be implemented efficiently using the algorithm *Lark Prune* (White 1991). The pseudocode of these two algorithms are provided in § 2.14).

Algorithm 1 Solving unconstrained POMDP

```
Solve POMDP ($\theta, \beta_0$)
 1: $\mathscr{A}_N \longleftarrow \big(\alpha_N = \langle V_N(s_1), ..., V_N(s_{|S|})\rangle, c_N = \vec{0}\big)$   ▷ each element in $\mathscr{A}_N$ is a tuple
 2: for $t$ in $\{N-1, ..., 0\}$ do   ▷ iterate over time backwards
 3:     for $a$ in $A_t$ do   ▷ iterate over actions
 4:         for $o$ in $O_{t+1}$ do   ▷ iterate over observations
 5:             for $v$ in $\mathscr{A}_{t+1}$ do   ▷ iterate over alpha-vector sets obtained from the last iteration
 6:                 $u \longleftarrow v[0], \chi \longleftarrow v[1]$ ▷ initialize $u, \chi$ to be the 1st and 2nd element of $v$, respectively
 7:                 $U \longleftarrow \varnothing$
 8:                 for $d$ in $[29]$ do   ▷ calculate the expected reward and cost in the next 29 days
 9:                     $u = \langle r_t(a, s_1), ..., r_t(a, s_{|S|})\rangle + \mathcal{P}_t^a \cdot u$
10:                     $\chi = \langle c_t^a(s_1), ..., c_t^a(s_{|S|})\rangle + \mathcal{P}_t^a \cdot \chi$
11:                 end for
12:                 $U.add((\tau(u, a, o), \tau_c(\chi, a, o)))$   ▷ calculate alpha-vector using Equations (2.19,2.23)
13:             end for
14:             $\mathscr{A}_t^{a,o} \longleftarrow$ Lark Prune($U$) ▷ whether $(u, \chi) \in U$ will be pruned depends on $u$, i.e., reward
15:         end for
16:         $\mathscr{A}_t^a \longleftarrow$ Incremental Pruning($\mathscr{A}_t^{a,o_1}, ..., \mathscr{A}_t^{a,o_{|O|}}$)   ▷ similar to Line 14
17:     end for
18:     $\mathscr{A}_t \longleftarrow$ Lark Prune $(\bigcup_a \mathscr{A}_t^a)$   ▷ similar to Line 14, and we keep each set $\mathscr{A}_t^a$ separate in $\mathscr{A}_t$
19: end for
20: $E \longleftarrow \arg\max\{\beta_0 \cdot (e[0]) : e \in \mathscr{A}_0\}$   ▷ find all elements in $\mathscr{A}_0$ with maximum expected reward
21: return $E$, $E$.action   ▷ return set $E$ and its associated action set $E$.action
```

Algorithm 2 Solving System II

```
Main Solver ($\Gamma_t, \beta_0$)
$\theta_0 \longleftarrow 0,\ \theta_1 \longleftarrow 1$    ▷ keep track of the lower bound and upper bound of the current $\theta$
$C_{old} \longleftarrow -\infty$    ▷ keeps track of the expected cost from the last iteration
$\Delta\theta,\ \Delta C \longleftarrow 1$    ▷ initialize the change in $\theta$ and the optimal expected cost $C$ to some number $> \epsilon$
$Y_0$, $Y_0$.action $\longleftarrow$ Solve POMDP($\theta = 0, \beta_0$)    ▷ check the extreme points
$Y_1$, $Y_1$.action $\longleftarrow$ Solve POMDP($\theta = 1, \beta_0$)
$m_0 \longleftarrow \arg\max_{y \in Y_0} y[1]$    ▷ find the feasible policy with lowest expected cost
$m_1 \longleftarrow \arg\min_{y \in Y_1} y[1]$    ▷ find the feasible policy with highest expected cost
**if** $|\beta_0 \cdot (m_0[1])| > \Gamma_t$ **then return** problem not feasible
**else if** $|\beta_0 \cdot (m_0[1])| = \Gamma_t$ **then return** $m_0$.action, $\beta_0 \cdot (m_0[0]), \beta_0 \cdot (m_0[1])$
**else if** $|\beta_0 \cdot (m_1[1])| \leq \Gamma_t$ **then return** $m_1$.action, $\beta_0 \cdot (m_1[0]), \beta_0 \cdot (m_1[1])$
**end if**
**while** $\Delta\theta > \epsilon$ or $\Delta C > \epsilon$ **do**    ▷ the loop stops if both $\Delta\theta \leq \epsilon$ and $\Delta C \leq \epsilon$
    $\theta \longleftarrow (\theta_0 + \theta_1)/2$    ▷ calculate current $\theta$ value
    $Y$, $Y$.action $\longleftarrow$ Solve POMDP($\theta, \beta_0$)    ▷ obtain the set of optimal solutions
    $m \longleftarrow \arg\max_{y \in Y} y[1]$    ▷ find the element with the lowest expected health
    **if** $|\beta_0 \cdot m[1]| = \Gamma_t$ **then** break loop    ▷ we have found the optimal $\theta^*$
    **else if** $|\beta_0 \cdot m[1]| < \Gamma_t$ **then** $\theta_0 \longleftarrow \theta$    ▷ if feasible, update the lower bound $\theta_0$ to the current $\theta$
    **else** $\theta_1 \longleftarrow \theta$    ▷ update the upper bound $\theta_1$ to $\theta$
    **end if**
    $\Delta\theta \longleftarrow (\theta_1 - \theta_0), \Delta C \longleftarrow |C_{old} - (m[1]) \cdot \beta_0|$    ▷ update the change in $\theta$ and $C$
    $C_{old} \longleftarrow (m[1]) \cdot \beta_0$    ▷ update $C_{old}$
**end while**
$\theta^*$, $Y^*$, $Y^*$.action $\longleftarrow \theta, Y, Y$.action    ▷ we have found $\theta^*$ in the above while loop
$m^* \longleftarrow \arg\min_{y \in Y^*} \{y[1] \ : \ y[1] \leq \Gamma_t\}$    ▷ solve Equation (2.8)
**return** $m^*$.action, $\beta_0 \cdot m[0], \beta_0 \cdot m[1], \theta^*$    ▷ return optimal action, expected reward, cost, and $\theta$
```

The Solution to Our Reformulation To compute the expected cost as defined in System II, similar to the definition of α_t, we let χ_t be a vector denoting the expected cost at each state s_t at time t. Thus, we have

$$\tau_c(\chi_{t+1}, a_t, o_{t+1})(s_t) = \frac{c_t^a}{|O|} + \sum_{s_{t+1} \in S} w(o_{t+1}|s_{t+1}, a_t)\mathcal{P}(s_{t+1}|s_t, a_t)\chi_{t+1}(s_{t+1}). \tag{2.23}$$

Let $\tau(u, a, o)$ and $\tau_c(\chi, a, o)$ be two α-*vector*s of length $|S|$, where the i^{th} element of $\tau(u, a, o)$ and $\tau_c(\chi, a, o)$ are $\tau(u, a, o)(s_i)$ and $\tau_c(\chi, a, o)(s_i)$, respectively. Algorithm 1 describes the complete algorithm to solve the unconstrained POMDP problem (Equation (2.11) in System II) aligning the transition and reward matrices with our one-month decision period (since the transition and reward matrices are defined in days). Note that the inputs of the three pruning steps, in Lines (14)–(18) of Algorithm 1, are sets of tuples with length 2, and whether a tuple will be pruned in those steps depends on the first entry of the tuple. That is, the three pruning steps are performed over the first entry of each element in the input sets.

In order to solve System II, one slight modification to the algorithm *Lark Prune* is necessary. In the original *Lark Prune* algorithm, if there are two policies that give the same optimal objective value, then the tie is broken arbitrarily. However, in our algorithm, we maintain all policies that produce the same optimal objective value and only discard an optimal policy if it admits the same expected health and cost with another policy.

Algorithm 1 includes only the necessary variables to demonstrate the solution methods of System II. For example, to obtain the expected health directly from the algorithm, one could add an additional element h_N towards the end of the tuple $\mathscr{A}_N$ in Line 1 and update the rest of the algorithm accordingly. Similarly, we could add variables to keep track of the expected health coming from the terminal reward. Furthermore, this structure allows us to plug in different transition matrices $\mathcal{P}_t^a$ in Lines (9)–(12) when comparing different models. We will discuss this in more detail in our numerical section, § 2.6.

Algorithm 2 solves our reformulation System II, where a binary search over the θ space is implemented to find the optimal θ^* value given the budget constraint and an initial belief state. To obtain randomized policies, one could discretize the space of Γ_t and find the convex hull of the optimal solution set. The complexity of our algorithm is exponential in the numbers of states and actions, and polynomial in the parameter ϵ. In the case where all states are fully observable, i.e., we have a MDP, Algorithm 2 can be modified to incorporate any exact solution methods for finite-horizon MDPs, and the complexity of our algorithm is polynomial in the numbers of states and actions, and ϵ.

Note that since the costs are always negative, the expected cost at a state s,

$$\chi_t(s) = \sum_{o \in O_{t+1}} \tau_c(\chi_{t+1}, a, o)$$

is non-increasing as t decreases in Algorithm 1 for all action a and state s. Thus, to reduce the run-time of our algorithm, in addition to pruning (Lines 14-18), we could remove an element in the set $\mathscr{A}_t$ if the expected cost evaluated at β_0 exceeds the budget. While our reformulation can be solved optimally within a reasonable time window when problem size is comparatively small (as in our problem), we provide a heuristic algorithm (with a worst-case error bound) in § 2.15 to further reduce the run-time.

2.6 Numerical results

In this section, we numerically investigate the benefit of incorporating wearable devices in OUD treatments using the reformulation and Algorithm 2 presented in § 2.5. We are collaborating with a startup[8] to test the feasibility of different wearables (Fitbit, Garmin, and Empatica E4) in capturing patients' craving states: thirty patients from Jade Wellness Center in Pennsylvania are participating in this pilot study. One initial observation is that craving frequencies differ greatly across patients[9], confirming the need to study the benefits of a personalized treatment strategy. However, since the data that we have collected thus far is insufficient for accurate parameter estimation, the majority of the parameters of our CPOMDP model were estimated from past literature. (We can confirm that these are in the range of the observed pilot data.) For those that were not available from the literature, we perform extensive sensitivity analysis. Our model is written in Python and takes an average of 60 (±25) seconds to run all 7 cases (Case 1, Case 2a, Case 2b, Case 2c, Case 2d, Case 3, and the benchmark case) to completion on a 3.2 GHz Core i7-8700 machine with 64 GB Ram. All cases took on average 14 iterations to terminate.

In § 2.6.1 we describe the various types of wearable devices. In § 2.6.2 we describe our transition matrices and parameter estimation. In § 2.6.3 we describe our numerical results. We discuss sensitivity analysis and case extensions in § 2.6.4. We denote the average transition probabilities of the entire patient population as the *average treatment dynamics* (ATD).

2.6.1 Comparison of Cases

Table 2.2 summarizes all cases that we study. Case 1 contains only monthly urine tests (cost \$360 per year), while Cases 2 and 3 additionally study three types of wearables (and two hypotheticals). The first type costs \$120 (Fitbit); the second type costs \$258 (Garmin); the third type costs \$1200

[8] We provided them with a first set of wearables. This startup subsequently received funding by the NSF to complete their SBIR Phase I study.

[9] Also observed in various past literature (Chalana et al. 2016). A more detailed discussion on factors that could affect patient treatment response can be found in § 2.2 and § 2.6.2.

Case Num.	Technology	Device Cost	Device Accuracy	Device Life Span
1	UT	-	-	-
2a	wearables with UT	\$120	0.6	1
2b	wearables with UT	\$258	0.7	1.5
2c	wearables with UT	\$500	0.8	2
2d	wearables with UT	\$800	0.9	2.5
3	wearables	\$1200	1	3
4	benchmark	0	1	-

Table 2.2: Costs associated with different cases. In Cases 1, 2, and 3, we assume the urine tests all cost \$360 per year. The device accuracy is measured in its sensitivity and specificity in detecting patients' craving episodes, and the device life spans are measured in years. UT stands for *urine test*.

(Empatica E4). Recall that in § 2.4, we defined the sensitivity (σ_s) and specificity (p_s) for every partially observable state, $s \in S$, for each type of wearable. Because our experimental results are robust under small perturbations in σ_s and p_s, we set $\sigma_s = p_s$ for all partially observable states s, and $\sigma_s = \sigma_{s'}$ for every partially observable states pair (s, s'). Henceforth, we refer to the value of σ_s associated with each wearable type as the device accuracy. According to the number of features that each type of wearable collects, we set the device accuracy to be 0.6, 0.7, and 1 for types 1, 2, and 3 wearables, respectively, where a device accuracy of 0.5 is equivalent to guessing randomly. To fill out intermediate values, we consider two additional hypothetical types of wearables: accuracy of 0.8 and cost \$500, and accuracy of 0.9 and cost \$800. Because costlier wearables have longer life spans, we spread the price over their life spans. We obtain our benchmark case, Case 4, by setting the cost of the wearable devices and urine test in Case 3 to zero.

Wearables can also help to determine patient response to different treatments (*personalized treatment dynamics*, PTD) in Cases 2, 3, and 4 (there is no information in Case 1 that will allow us to do that). In our experiments, we compare the scenarios where PTD is equal to or divergent from ATD. In addition, we consider three levels of patient treatment adherence levels, where high, medium, and low TAs represent that the patient follows the treatment above 90%, between 70% and 90%, and below 70% of the time, respectively.

2.6.2 Transition Probability and Parameter Estimation

Transition Probability Matrices Recall that $\mathcal{P}(s_{t+1}|s_t, a)$ is the transition probability from state $s_t \in S$ to state $s_{t+1} \in S$ under action $a \in A$, where t indicates the number of days since the last drug-use. Let $\mathcal{P}_t^a$ be the daily transition probability matrix in which the rows correspond to s_t

and columns to s_{t+1}. Since patients' treatment adherence levels can affect the evolution of their conditions and hence the transition probabilities, having different treatment adherence levels could potentially affect the additional health benefits brought by wearable devices. Thus, $\mathcal{P}_t^a$ is also a function of a patient's treatment adherence level.

Although past literature also suggests that other variables—including the patient's age, jail history, comorbid medical conditions, drug injection history, number of past overdoses, and number of relapses—could also affect the transition probabilities, we assume that theses variables are automatically captured in the resulting transition matrices. In addition, since the horizon in our model is short (at most twelve months), we assume that the numbers of past overdoses and relapses remain the same (regardless of the possibility that the patient might reach states OD or Re in the rest of the horizon), to ensure the tractability of our model. Since the budget is modeled as an open-loop optimization problem (i.e. at the beginning of each month, we solve a new optimization problem with an updated budget and the patient's information), if the patient relapsed within a month, then the transition matrices will first be updated in the following month to incorporate this information.

Omitting t and a, let x_s, nc_s, c_s^2, e_s, and od_s be the probability of transitioning to Dx, NC, C2, Re, and OD, from health state s, respectively; let d_s be the probability of dying at state s and w be the probability of withdrawing from any state with any action at any time.[10] We represent the transition probability matrices as follows:

$$\mathcal{P}_t^a = \begin{array}{c} \\ \end{array}\begin{array}{c|cccccccc|}
\textit{from/to} & \text{Dx} & \text{NC} & \text{C1} & \text{C2} & \text{Re} & \text{OD} & \text{Dt} & \text{Abs} \\
\hline
\text{Dx} & x_{\text{Dx}} & \text{nc}_{\text{Dx}} & Z_1 & c_{\text{Dx}}^2 & 0 & \text{od}_{\text{Dx}} & d_{\text{Dx}} & w \\
\text{NC} & 0 & \text{nc}_{\text{NC}} & Z_2 & 0 & 0 & 0 & d_{NC} & w \\
\text{C1} & 0 & \text{nc}_{\text{C1}} & Z_3 & c_{\text{C1}}^2 & e_{\text{C1}} & \text{od}_{\text{C1}} & d_{\text{C1}} & w \\
\text{C2} & 0 & 0 & Z_4 & c_{\text{C2}}^2 & e_{\text{C2}} & \text{od}_{\text{C2}} & d_{\text{C2}} & w \\
\text{Re} & x_{\text{Re}} & 0 & Z_5 & c_{\text{Re}}^2 & e_{\text{Re}} & \text{od}_{\text{Re}} & d_{\text{Re}} & w \\
\text{OD} & 1 - d_{\text{OD}} & 0 & 0 & 0 & 0 & 0 & d_{\text{OD}} & 0 \\
\text{Dt} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\text{Abs} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array},$$

where $Z_1 = 1 - x_{\text{Dx}} - \text{nc}_{\text{Dx}} - c_{\text{Dx}}^2 - \text{od}_{\text{Dx}} - d_{\text{Dx}} - w$, $Z_2 = 1 - \text{nc}_{\text{NC}} - d_{\text{NC}} - w$, $Z_3 = 1 - \text{nc}_{\text{C1}} - c_{\text{C1}}^2 - e_{\text{C1}} - \text{od}_{\text{C1}} - d_{\text{C1}} - w$, $Z_4 = 1 - c_{\text{C2}}^2 - e_{\text{C2}} - \text{od}_{\text{C2}} - d_{\text{C2}} - w$, $Z_5 = 1 - x_{\text{Re}} - c_{\text{Re}}^2 - e_{\text{Re}} - \text{od}_{\text{Re}} - d_{\text{Re}} - w$.

Throughout our paper, we assume that if a patient reaches any of the states Dx, OD, Dt, or Abs, clinicians will be notified immediately; that is, these states are always observable. Fig. 2.2 depicts the transition diagram between our health states.

[10] The assumption that w is independent of states and action is observed by Termorshuizen et al. (2005).

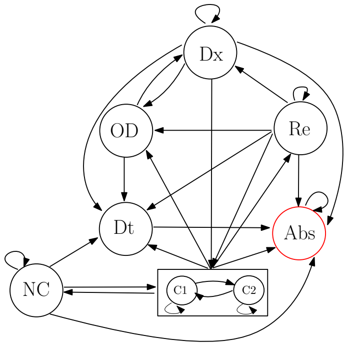


Figure 2.2: Transition diagram. States C1 and C2 share the same types of transitions, however the transition probabilities need not to be the same.

According to Chalana et al. (2016), if patients relapse right after detoxification, they will overdose with high probability due to the loss of tolerance. After experimenting with various values, we find that our results are robust with respect to the probability transitioning from Dx to Re, e_{Dx}, and thus we set it to be 0 (i.e. a patient will overdose with probability 1 if relapsing the day after detoxification). Moreover, in our transition matrices, the probability of transitioning to state OD decreases as t increases. We assume that patients either die or receive treatments followed by detoxification if they overdose, and that patients must experience some craving the day before relapse if they are not in state Dx.

Parameter Estimation We initialize two sets of transition matrices $\{T_0^a\}_{a\in A}$ and $\{T_1^a\}_{a\in A}$. The former is estimated from past literature (§ 2.16) and reflects how a patient reacts to different treatments on average, i.e., the ATD. The latter matrix is randomly perturbed around the ATD to represent the true treatment dynamics of a given patient (patient ground truth). Therefore, to find the optimal policy, we set the transition matrices to be $\{T_0^a\}_{a\in A}$ in Case 1 (since we do not have access to wearables and thus could not develop PTD), and set them to be $\{T_1^a\}_{a\in A}$ in Cases 3 and 4. In Case 2, since the wearables provide imperfect information, we assume that the accuracy of the estimated transition probability matrices increases as the accuracy of the wearable increases. Thus, we randomly perturb the transition matrices $\{T_1^a\}_{a\in A}$ in Case 2 with the perturbation magnitude decreasing as the wearable accuracy increases[11]. We refer to the resulting expected health and cost as the *estimated (or observed) expected health and cost respectively.*

To evaluate and compare the expected health in each case, we use $\{T_1^a\}_{a\in A}$ as our transition probability matrices and the identity matrix as our observation matrix to calculate the expected health, and we refer to these as the *true expected health* and *cost* respectively. Past literature shows that the treatment IN is superior to treatment M and B; therefore, we assume in our experiment

[11] Thus the information our algorithm uses is a more accurate estimate of the true transition matrices.

that $\mathcal{P}_t^{\mathrm{IN}}$ is stochastically equal to or better than $\mathcal{P}_t^{\mathrm{M}}$ and $\mathcal{P}_t^{\mathrm{B}}$. The effects of TA are estimated from Nosyk et al. (2009). In the rest of this section, unless otherwise stated, the *expected health* refers to the *true expected health.*

In our experiments, we consider a patient with the following characteristics: the patient has relapsed 5 times, overdosed once, and has just completed a 21-day detoxification program. (We experimented with different patient characteristics, and the results are similar.) We assume that the initial belief state is high craving intensity (C2), and that an average patient reacts to treatment B better than to treatment M. To avoid overly myopic solutions, we calculate the expected health gain for the patient under no treatment from day 360 to day 420, weight it by θ, and add it as the terminal reward.

2.6.3 Results

In our baseline scenario (§ 2.6.3), we examine the benefit of adopting wearables in OUD treatments in the case where the patient reacts to treatments exactly as the average treatment dynamics predict, implying that, by definition, there is no benefit of implementing PTD. We consider three budget levels: low ($9K), medium ($15K), and high ($21K). When the budget is low, we could not afford MAT in every single period; when the budget is medium, we could afford MAT is all periods and can afford to treatment IN occasionally; when the budget is high, we could afford treatment IN in the majority of periods. Finally, we explore two possibilities within PTD: when the patient reacts to treatment M better than to B (§ 2.6.3) and when the patient reacts to treatment B better than to M (§ 2.6.3).

Since all cases contain monthly urine tests, we assume the state relapse is observable (at the end of the month). Thus, in this section, we have only three partially observable states: NC, C1, and C2, and all cases are solved using the exact algorithm. Unless stated otherwise, our results are robust with respect to terminal rewards. Because our budget formulation finds only the points that lie on the efficiency frontier of the health-budget curve, and we consider only deterministic optimal policies in our numerical experiments below (as randomized policies may be problematic in the field), we sometimes observe that our optimal solution does not spend all the budget. The *cost gap*, the gap between the observed expected cost and our budget, indicates how close our solution is to the optimal solution when the problem is fully observable. We define the *value*s of wearables to be the difference between the expected health when we incorporate wearables and that of Case 1.

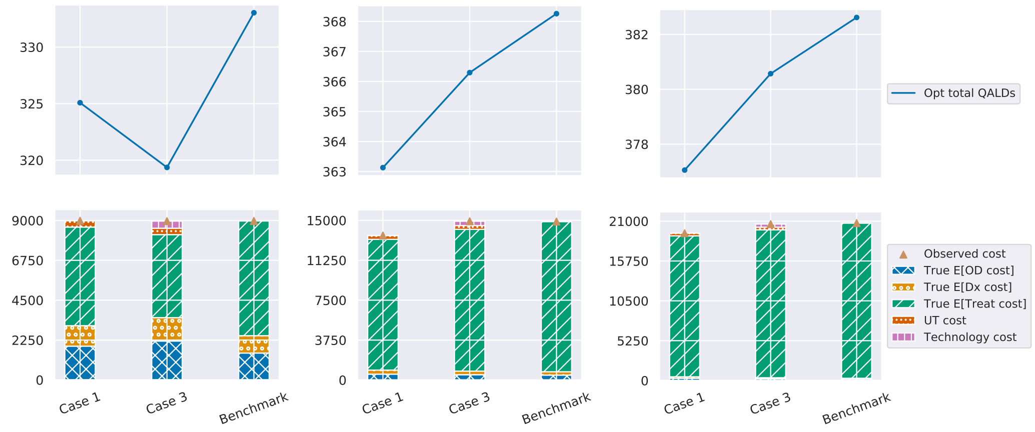


Figure 2.3: The Baseline Scenario. Left: 9K budget with medium TA. Middle: 15K budget with medium TA. Right: 21K budget with medium TA. At termination, the optimal θ's are approximately [0.053, 0.039, 0.077] (left), [0.294, 0.269, 0.274] (middle), and [0.301, 0.303, 0.310] (right), where the number of digits that θ^* contains equals to the number of iterations that took System II to terminate; the expected QALDs received from treatments (excluding the terminal reward) are [285.44, 280.98, 289.58] (left), [313.0, 316.03, 317.9] (middle), and [326.18, 329.49, 311.36] (right).

Baseline scenario

In this scenario, the patient reacts to different treatments in keeping with the average treatment dynamics. Fig. 2.3 (left top) shows the true expected QALD gains under the optimal policy for different cases. In Fig. 2.3 (left bottom), the estimated expected cost is represented by the triangle; the true expected costs of different elements are represented by the colored bars. Since the improvement in health between the benchmark model and Case 1 is limited, in Fig. 2.3 we omit Case 2 for ease of illustration. As we will see later, the values of Case 2 wearables are often dominated by that of Case 3.

When the budget is \$9K (Fig. 2.3 left), the value of the wearable is negative. This is because the benefit of observing health states with higher certainty is undermined by the fact that we are left with less money to treat the patient. As the budget increases to \$15K and \$21K (Fig 2.3 middle and right), the value of the wearables is positive but remains negligible: the health state relapse is observable and the rewards for different craving states are similar, so the benefits of wearables are limited. Further experiments suggest that this observation holds for patients with different TA levels and different transition matrices. In addition, we observe in Fig. 2.3 that for each case, as the budget increases, the optimal θ value indeed increases as indicated by Proposition 2(b). In this section, we conclude that if all patients react to treatments in accordance with the average treatment dynamics—that is, individual variability is low, then we should not adopt wearables in OUD treatments.

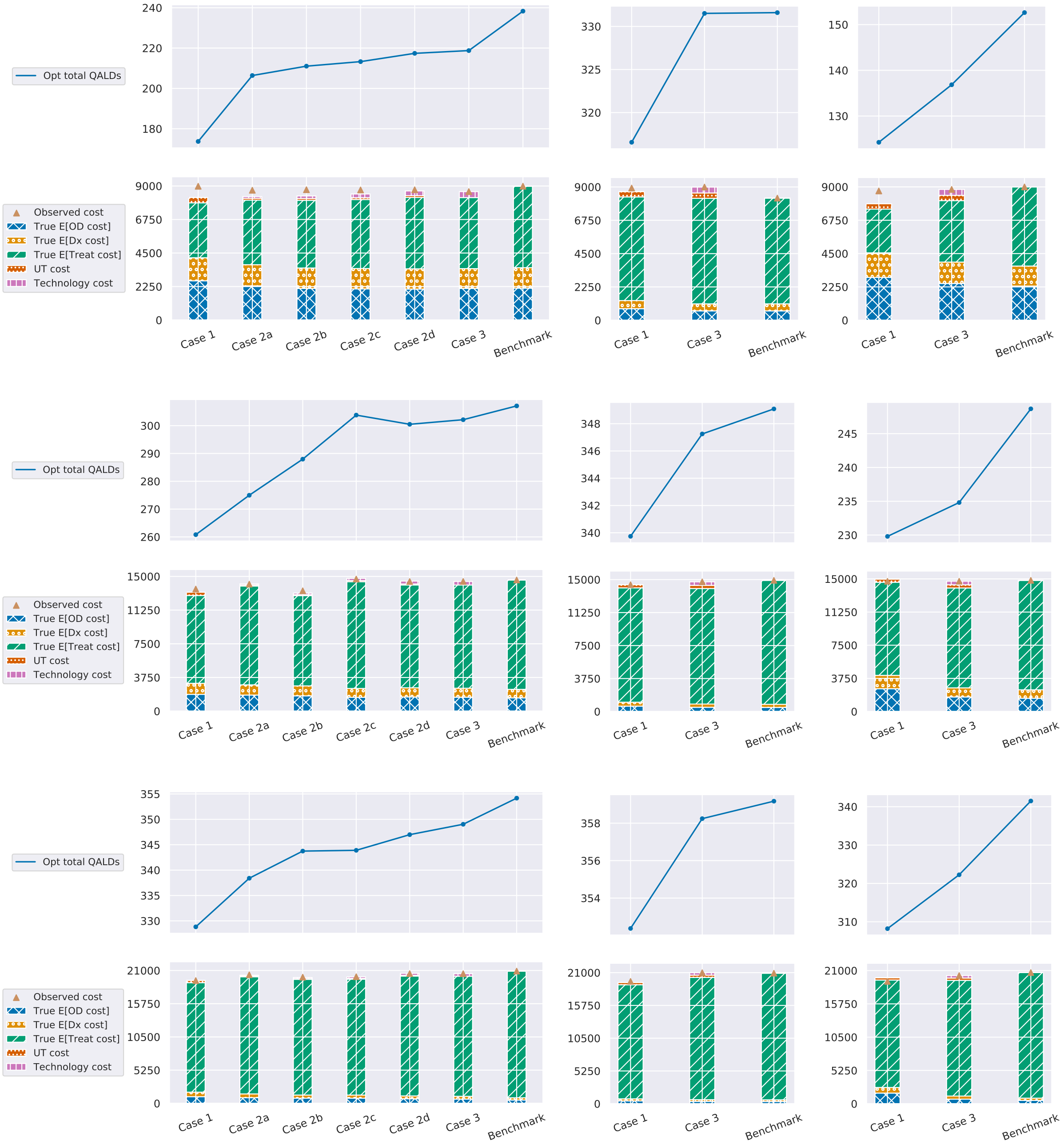


Figure 2.4: Scenario 1 where the patient reacts to treatment M better than B. Top row: 9K budget. Middle row: 15K budget. Bottom row: 21K budget. Left column: medium TA. Middle column: high TA. Right column: low TA. At termination, from top left to bottom right, the optimal θ's are approximately [0.052, 0.039, 0.030, 0.043, 0.037, 0.036, 0.040], [0.406, 0.115, 0.205], [0.099, 0.055, 0.056], [0.294, 0.187, 0.135, 0.121, 0.106, 0.102, 0.103], [0.534, 0.335, 0.339], [0.126, 0.058, 0.058], [0.300, 0.222, 0.181, 0.165, 0.145, 0.133, 0.141], [0.551, 0.419, 0.432], and [0.146, 0.060, 0.060].

Scenario 1

In this scenario, the patient reacts to treatment M better than to B, and PTD is implemented in Cases 2, 3, and 4. We first assume that the patient is allocated with a $9K budget and has medium TA. We observe in Fig. 2.4 (top left) that the expected health increases as the accuracy of the wearable device increases. In addition, the cost gap decreases significantly as the device accuracy increases. In Fig. 2.4 (top middle) we observe that the value of wearable devices decreases significantly. This is because the high adherence causes the outcome of less expensive treatment (for example, treatment M) to become comparable to that of the most expensive treatment, treatment IN. Similarly, Fig. 2.4 (top right) shows that the value of wearables decreases when comparing to Fig. 2.4 (top middle), this time because patients with poor treatment adherence react poorly to both treatments M and B. When comparing figures in Fig. 2.4 top row, we observe that when the budget is low, the expected costs of overdose and detoxification increases as the patient treatment adherence level decreases. This is because patients with lower adherence levels are more likely to relapse when the treatment is inconsistent, i.e., when we cannot afford to provide treatment in every single period.

Now, if we increase the budget from $9K to $15K, in Fig. 2.4 (middle left), we find that the value of more expensive wearables increases slightly when compared with Fig. 2.4 (top left) Although the values of wearables remain positive when budget is increased to $21K in Fig. 2.4 (bottom left) they are reduced as the more expensive treatment is utilized more frequently. Again, when TA is high or low, we observe in Fig 2.4 (middle middle and right, bottom middle and left) that the value of the wearable is less than that when TA is medium. Moreover, the value of the wearables slightly decreases as the budget increases because we can afford treatment IN more frequently. Overall, the value of wearables increases when patients' PTD vary from the ATD.

Scenario 2

In this scenario, the patient reacts to different treatments in the same order as the ATD (i.e., reacts to treatment B better than M), but with a different magnitude. Because many patterns we observe in this scenario are similar to those in Scenario 1, we highlight the differences. We observe that in Fig. 2.5 (middle), the value of the wearable becomes negative when compared with Fig. 2.4 (top middle). In Fig. 2.5 (left) because the cost of the urine test and wearable device is high in Case 3 when compared to our budget, although the increase in treatment outcome between Cases 4 and 1 is relatively large, the health improvement is relative minor in Case 3. In Fig. 2.5 (right) when the patient has low treatment adherence, we observe that the value of wearables increases slightly when compared with that of Scenario 1 in Fig. 2.4 (top right). In addition, this value is similar to the one observed in Fig. 2.5 (top left) When we increase the budget in Scenario 2, we observe in Fig. 2.7 (§ 2.17) that wearables can still improve patient treatment outcomes. Finally, in § 2.6.3 and

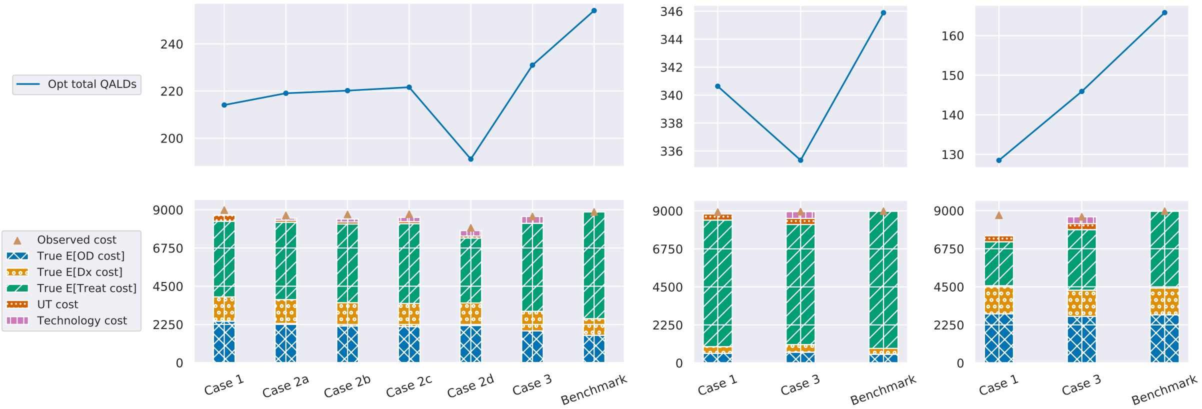


Figure 2.5: Scenario 2 where the patient reacts to treatment B better than M but with a different magnitude than in the average treatment dynamics. Budget level equals to 9K. TA equals to medium, high, and low from left to right. At termination, from left to right, the optimal θ's are approximately [0.052, 0.034, 0.025, 0.025, 0.025, 0.027, 0.029], [0.406, 0.026, 0.106], and [0.099, 0.052, 0.053]; the expected QALDs received from treatments (excluding the terminal reward) at termination are [200.73, 202.68, 200.19, 200.97, 172.13, 206.40, 224.61], [297.29, 293.30, 301.72], and [125.33, 135.47, 152.74].

§ 2.6.3, we observe that when the accuracy of the wearables is low, we often observe a large gap between the observed and actual expected cost and thus either undertreat the patients or exceed the budget constraint.

2.6.4 Sensitivity Analysis and Case Extensions

Sensitivity to the Cost of Urine Tests Recall that the cost of urine tests is $360 per year, which is higher than the cost of a wearable with device accuracy less than 0.8. On the other hand, if the wearable has high device accuracy, then monthly urine tests become less valuable. Thus, in our experiments, we also explored the possibility of having less frequent urine tests for Case 2 (while the state relapse remains fully observed). Similar to § 2.6.3 and § 2.6.3, we observe wearables with low to medium device accuracies often underestimate the costs (e.g., in Fig. 2.8 in § 2.17). When we compare Fig. 2.8 with Fig.s 2.4, 2.5, and 2.7, we find that having less frequent urine tests often increases the health outcome even when the device accuracy is low, as the money can be better spent. However, under different matrix perturbation schemes, we sometimes observe that having more frequent urine tests is beneficial when device accuracy is low.[12] In contrast, having less

[12] Recall that the actions in our model can only be changed once per month, and we assume that the urine test result is obtained right before considering the change of treatment in the next period.

frequent urine tests always increases the health outcome when device accuracy is high.

Sensitivity to Other Model Parameters To check the robustness of our results, we experimented with different mechanisms to perturb the transition probability matrices and observe similar results with a few exceptions: 1) the magnitude of the value of wearables when the patient has low treatment adherence and reacts to treatment M better than B is sensitive to different matrix perturbation mechanisms (e.g., see Fig. 2.9 (left) in § 2.17); 2) when the device accuracy is medium, we sometimes observe that the value of wearables could also be negative (e.g., see Fig. 2.9 (right) in § 2.17) because the increase in the cost of the wearables exceeds the benefit of the additional information gathered by the wearables. Both observations highlight the importance of knowing the exact device cost-accuracy tradeoffs, and could be potential research topics for future field studies.

We found that the rest of our insights are robust under small perturbations in the transition matrices, as long as the stochastic order of patient reaction to different treatments remains the same. This robustness also holds for immediate rewards, costs, and the initial belief state. However, our model can be sensitive to the terminal rewards. In particular, if the terminal reward is 0 and the budget is low, our optimal solution becomes myopic, in the sense that the patient might end up in bad health states in the end of the horizon with high probability. On the other hand, one should also be cautious when setting a high terminal reward, as the optimal policy might be changed: intuitively, when the terminal reward is too large, we will treat patients with the most expensive treatment towards the end of the horizon to try to gather the terminal reward.

Extensions We extended our cases to incorporate EMA survey devices where daily survey responses are collected from patients. In particular, we consider the possibility that the patient may provide truthful, random, or falsified information. We describe the model for this extension in § 2.18. In this case the transition probability matrices could be directly estimated from the survey results. Thus, the accuracy of the transition probability matrices would decrease when the patient provides noisy or falsified information more frequently. While the value of EMA survey devices dominates that of wearables when completion frequency is high and the patient is completely honest, we see that in our field study that patients or their surrogates do not answer surveys very regularly, thus reducing the feasibility of EMA devices. At the extreme, the value of EMA device can be negative when the patient provides only noisy information.

2.7 Summary of Findings and Concluding Remarks

In this work, we have built a POMDP model with budget constraints to evaluate the potential value of incorporating different wearables in OUD treatments. We provide a novel budget reformulation and show that it can be solved efficiently.

Our experiments with our model, incorporating parameters calibrated to literature, show that value of wearables increases as the patient treatment dynamics deviate from the ATD. In addition, as the budget increases, the value of wearable devices eventually decreases when compared to our baseline model. This is because when the budget is relatively low, choosing the cost-effective treatment could potentially increase patient treatment outcome by a large margin. When the budget is sufficiently high, treatment IN guarantees the same treatment outcome for all patient treatment adherence groups and is superior to all other treatments. The value of wearables thus becomes negligible.

We also discover that wearables are more beneficial to patients with medium to low treatment adherence levels; patients with high treatment adherence react to cheaper treatments better than those with lower treatment adherence, and it becomes more cost-effective to treat the patient even without the wearables. Similarly, we observe that when the budget is fixed, as the patient treatment adherence decreases, the patient is more prone to relapse; thus, incorporating wearables can effectively reduce the risk of relapse. Not surprisingly, as the budget increases, the marginal return of money spent on treatments or wearables decreases the fastest for patients with high treatment adherence.

These observations imply that when the budget that we have is large enough to treat every patient with treatment IN, there is no benefit of adopting wearables. On the other hand, if the budget is smaller, to maximize the treatment outcome for the entire patient group, we should prioritize patients with higher treatment adherence; in this case, the benefit of incorporating wearables in OUD treatments is also limited. However, in the situation where we have enough budget to treat every patient but cannot afford treatment IN for everyone, to maximize the treatment outcome for the entire patient group, we should allocate more money to patients with lower treatment adherence and incorporate wearables for those patients. Although maximizing treatment outcome for the entire patient population is a reasonable objective, the ethics behind which patient we should allocate resources to is debatable. This question, however, is beyond the scope of this paper.

Future work includes collecting data from planned field studies in collaboration with rehab centers and start-ups that are developing wearable devices for tackling OUD to fine tune our models and analysis. Throughout the paper, we have assumed that the use of wearables in OUD treatments does not affect patient TA levels as well as treatment outcomes. It would also be interesting to see if there are behavioral changes attributable to wearables, even if we only passively collect data, i.e., not sending warnings/reminders, as not all types of wearables have this capability. We hope our work here adds to the growing literature and motivates further work to tackle this important issue.

2.8 Case 1 Parameter Calculation

Let 30C1 be the event of experiencing C1 on day 30. Then

$$c1_{ut+} = P(ut+|30\text{C1}) = \frac{P(ut+\cap 30\text{C1})}{P(30\text{C1})} = \frac{P(ut+\cap 30\text{C1})}{\sum\limits_{i=1}^{8}((f_t^a)^{30})_{i,3}/8},$$

where $((f_t^a)^{30})_{i,3}$ is the i^{th} row and 3^{rd} entry of $(f_t^a)^{30}$. Because all rows of $(f_t^a)^{30})_{i,3}$ sum up to 8, which is the number of our states, we renormalize the value of $P(\text{C1 on day 30})$ such that it is between 0 and 1. Furthermore,

$$P(ut+\cap 30\text{C1}) = P(30\text{C1}|\ 27\text{Re})P(27\text{Re}) + P(30\text{C1}|28\text{Re})P(28\text{Re}) + P(30\text{C1}|29\text{Re})P(29\text{Re}),$$

where 27Re, 28Re, and 29Re are defined the same way as 30C1, that is, experiencing relapse on day 27, 28, and 29, respectively. For example, $P(30\text{C1}|27\text{Re}) = ((f_t^a)^3)_{5,3}$ and $P(27\text{Re}) = \sum\limits_{i=1}^{8}((f_t^a)^{27})_{i,5}/8$.

2.9 Values of Immediate Rewards and Costs

States	Health
Dx	0.6*
NC	1 †
C1	0.9
C2	0.75
Re	0.683
OD	0.1*
Dt	0

Abs	withdraw the program from state Dx:	0.5 × remaining horizon*
	from state NC:	1 × remaining horizon
	from state C1:	0.9 × remaining horizon
	from state C2:	0.75 × remaining horizon
	from state Re:	0.678 × remaining horizon
	from state Dx:	0.5 × remaining horizon

Table 2.3: Health reward associated with each state. *: assumptions that we made based on conversations with clinicians. In sensitivity analysis, we found that our results are robust under small perturbations of those parameters. The rest of the values are generated according to Connock et al. (2007). †: the health reward of state NC is the baseline utility.

Item	Cost type	Costs
Urine tests	model cost	\$30/month (Schackman et al. 2012)
Wearable devices	model cost	see Table 2.2
EMA survey device messaging cost	model cost	\$60/year (Model in Appendix 2.18)
Physician and nursing time	treatment cost	\$50/month (Schackman et al. 2012)
Methadone maintenance treatment	treatment cost	\$20/month (Barnett 2009)
Buprenorphine maintenance treatment	treatment cost	\$200/month (Barnett 2009)
Implant Naltrexone (Vivitrol)	treatment cost	\$1200/month
Counseling	treatment cost	\$600/month
OD	state cost	\$2500/episode*
Detoxification	state cost	\$1500/episode

Table 2.4: The costs considered in the model. The values without references in the table were determined through conversations with clinicians from the perspective of the healthcare system. *: the expected cost of overdose includes the expected cost of hospitalization and emergency room visit.

Action/Treatment	Costs
No treatment	0
Methadone maintenance treatment with counseling	\$670/month
Buprenorphine maintenance treatment with counseling	\$850/month

Implant Naltrexone with counseling	\$1800/month
Counseling	\$600/month

Table 2.5: Action associated costs. This table summarizes the cost of taking each action using the values provided in Table 2.4.

2.10 Proof of Lemma 2

Lemma 2. *If the immediate reward and terminal reward of a discrete time finite horizon POMDP takes the form $\theta h_t^a(\beta_t) + (1-\theta)c_t^a(\beta_t)$ for every time step, then $V^\pi(\beta_0, \theta)$ can be written as $V^\pi(\beta_0, \theta) = \theta H^\pi_{\beta_0} + (1-\theta)C^\pi_{\beta_0}$.*

Proof. Proof of Lemma 2 Let N be the length of the horizon. We will proceed by induction, and show that at each step t, the value function of an action a_t and belief β_t, $V_t^{a_t}(\beta_t, \theta)$, can be transformed into

$$V_t^{a_t}(\beta_t, \theta) = \theta \hat{h}_t^a(\beta_t) + (1-\theta)\hat{c}_t^a(\beta_t). \tag{2.24}$$

First, recall that the value function of an action a with belief β_t of a general POMDP is

$$V_t^a(\beta_t) = \sum_{s_t \in S} r_t(s_t, a)\beta_t(s_t) + \sum_{o_{t+1} \in O} \mathcal{P}(o_{t+1}|a, \beta_t) V_{t+1}^*(\beta_{t+1})$$

Note first that our immediate reward at time t under treatment a at state s has the form $r_t(a,s) = \theta h_t^a(s) + (1-\theta)c_t^a(s)$, where $\theta \in [0,1]$, $h_t^a(s) \in [0,1]$ and $c_t^a(s) < 0$, for $t \in \{0, ..., N-1\}$, and our terminal reward at state s at time N is $r_N(a,s) = \theta h_N(s) + (1-\theta)c_N(s)$.[13]

Base case: Equation (2.24) holds when $t = N-1$. First, at $t = N$ for a fixed θ, the value function is

$$V_N(\beta_N, \theta) = \sum_{s_N \in S} \theta h_N(s)\beta_N(s) + (1-\theta)c_N(s)\beta_N(s) = \theta h_N(\beta_N) + (1-\theta)c_N(\beta_N).$$

Then, at $t = N-1$ with action a and belief β_{N-1}, the value function of our POMDP satisfies

$$\begin{aligned} V_{N-1}^a(\beta_{N-1}, \theta) &= \sum_{s \in S_{N-1}} r_{N-1}(s,a)\beta_{N-1}(s) + \sum_{o \in O_N} \mathcal{P}(o|a, \beta_{N-1}) V_N(\beta_N, \theta) \\ &= \theta h_{N-1}^a(\beta_{N-1}) + (1-\theta)c_{N-1}^a(\beta_{N-1}) + \sum_{o \in O_N} \mathcal{P}(o|a,\beta_{N-1}) \left(\theta h_N(\beta_N) + (1-\theta)c_N(\beta_N)\right) \\ &= \theta H_{N-1}^a(\beta_{N-1}) + (1-\theta)C_{N-1}^a(\beta_{N-1}), \end{aligned}$$

[13] Note that in our problem, $c_N(s) = 0$, but we want to prove our theorem under the most general setup.

where $H^a_{N-1}(\beta_{N-1})$ and $C^a_{N-1}(\beta_{N-1})$ are the expected health and cost at time $N-1$ under action a and belief β_{N-1}, respectively:

$$H^a_{N-1}(\beta_{N-1}) = h^a_{N-1}(\beta_{N-1}) + \sum_{o\in O_N} \mathcal{P}(o|a,\beta_{N-1})h_N(\beta_N), \text{ and}$$
$$C^a_{N-1}(\beta_{N-1}) = c^a_{N-1}(\beta_{N-1}) + \sum_{o\in O_N} \mathcal{P}(o|a,\beta_{N-1})c_N(\beta_N).$$

Induction step: to show Equation (2.24) holds for any step t, we assume it holds for step $t+1$, and show that it also holds for step t:

$$\begin{aligned} V^a_t(\beta_t,\theta) &= \sum_{s\in S} r_t(a,s)\beta_t(s) + \sum_{o\in O_{t+1}} P(o|a,\beta_t)V^{a_{t+1}}_{t+1}(\beta_{t+1},\theta) \\ &= \theta h^a_t(\beta_t) + (1-\theta)c^a_t(\beta_t) + \sum_{o\in O_{t+1}} P(o|a,\beta_t)\left(\theta H^{a_{t+1}}_{t+1}(\beta_{t+1}) + (1-\theta)C^{a_{t+1}}_{t+1}(\beta_{t+1})\right) \\ &= \theta H^a_t(\beta_t) + (1-\theta)C^a_t(\beta_t), \end{aligned}$$

where $H^a_t(\beta_t)$ and $C^a_t(\beta_t)$ are the expected health and cost at time t under action a and belief β_t, respectively:

$$H^a_t(\beta_t) = h^a_t(\beta_t) + \sum_{o\in O_t} \mathcal{P}(o|a,\beta_t)H^{a_{t+1}}_{t+1}(\beta_{t+1}), \text{ and}$$
$$C^a_t(\beta_t) = c^a_t(\beta_t) + \sum_{o\in O_t} \mathcal{P}(o|a,\beta_t)c^{a_{t+1}}_{t+1}(\beta_{t+1}).$$

■ □

2.11 The Remaining Proofs of Proposition 2

Let's first recall Proposition 2:

Proposition 2. *Properties of the solution set in System II:*

2*(a) The optimal policy $\pi^*(\Gamma_t)$ is not necessarily unique, but the values $H^{\pi^*}_{\beta_0}(\Gamma_t), C^{\pi^*}_{\beta_0}(\Gamma_t)$ are unique for any fixed Γ_t.*

2*(b) $H^{\pi^*_\theta}_{\beta_0}$, $-C^{\pi^*_\theta}_{\beta_0}$, and the objective function $\theta H^{\pi^*_\theta}_{\beta_0} + (1-\theta)C^{\pi^*_\theta}_{\beta_0}$ are all non-decreasing in θ. Furthermore, as Γ_t increases, θ^* is non-decreasing.*

2*(c) $H^{\pi^*}_{\beta_0}(\Gamma_t)$ is non-decreasing in Γ_t.*

2*(d) If $\gamma_2 > \gamma_1$ and $V^{\pi^*}(\beta_0,\gamma_2) > V^{\pi^*}(\beta_0,\gamma_1)$, $H^{\pi^*_{\theta^*}}_{\beta_0}(\gamma_2) > H^{\pi^*}_{\beta_0}(\gamma_1)$.*

2*(e) Let $\mathcal{H}^{\pi^*}_{\beta_0}$ denote the unique elements contained in the set $\left\{H^{\pi^*}_{\beta_0}(\Gamma_t)\right\}_{\Gamma_t\in(0,\infty)}$. Then, the set $\mathcal{H}^{\pi^*}_{\beta_0}$ is finite.*

2.11.1 Proof of Property 2(b)

Recall by Lemma 2, at each time t, for a fixed belief β_t, the optimal Bellman's equation for our POMDP (Equation (2.11)) at time t and belief state β_t can be written as a convex combination of the expected health and expected cost:

$$
\begin{aligned}
V_t^*(\beta_t, \theta) &= \max_{a \in A_t} \left\{ \theta h_t^a(\beta_t) + (1-\theta)c_t^a(\beta_t) + \sum_{o \in O_{t+1}} \mathcal{P}(o|a, \beta_t) V_{t+1}^*(\beta_{t+1}) \right\} \\
&= \max_{a \in A_t} \left\{ \theta H_t^a(\beta_t) + (1-\theta)C_t^a(\beta_t) \right\},
\end{aligned}
$$

where $H_t^a(\beta_t)$ and $C_t^a(\beta_t)$ are the expected health and cost at time t under action a and belief β_t (as defined in § 2.10). Let $V_t^a(\beta_t, \theta) = \theta H_t^a(\beta_t) + (1-\theta)C_t^a(\beta_t)$. Note that $H_t^a(\beta_t)$ and $C_t^a(\beta_t)$ are fixed under the optimal value function from the last iteration, $V_{t+1}^*(\beta_{t+1})$, current belief β_t, and action a, where $V_{t+1}^*(\beta_{t+1})$ is treated as a constant in the the optimal Bellman's equation. To ease notation, for a fixed β_{N-1}, let $a^*(\theta)$ be the action that yields the highest expected reward at time $N-1$ for some fixed parameter θ, that has the lowest expected health:

$$
a^*(\theta) = \operatorname*{arg\,min}_{a' \in \arg\max_{a \in A_{N-1}} V_{N-1}^a(\beta_{N-1}, \theta)} H_{N-1}^{a'}(\beta_{N-1}).
$$

It suffices to show that at time $t = N-1$, for every fixed β_{N-1}, the first part of Property 2(b) holds, that is,

Claim 1. *as θ increases, $H_{N-1}^{a^*(\theta)}(\beta_{N-1}, \theta)$ $-C_{N-1}^{a^*(\theta)}(\beta_{N-1}, \theta)$, and $V_{N-1}^{a^*(\theta)}(\beta_{N-1}, \theta)$ are non-decreasing in θ.*

This is because 1) the optimal Bellman's equation at any time step t can be written in the same format as $t = N-1$, and 2) the exact backward induction algorithms in solving POMDP relies on fixing the belief state at every searching point. (See § 2.5.3 for detailed discussion on one of the exact POMDP solution method)

Before proving Claim 1, we first introduce some notation and a key concept in the proof—*the effective action set*. Since we fix the belief β_{N-1} for the rest of the proof, we abbreviate $H_{N-1}^{a^i}(\beta_{N-1})$ with H_{a^i}, abbreviate $C_{N-1}^{a^i}(\beta_{N-1})$ with C_{a^i}, and abbreviate $V_{N-1}^a(\beta_{N-1}, \theta)$ with $V(a, \theta)$. Under a fixed belief state β_{N-1}, given a set of actions at time $N-1$, A_{N-1}, we define the effective action set $A_{\beta_{N-1}}$ to be the largest subset of A_{N-1} such that all actions lie on the efficiency frontier of the health-cost curve. That is, for every pair of actions in a^1 and a^2 in $A_{\beta_{N-1}}$, if $H_{a^1} < H_{a^2}$, then $-C_{a^1} < -C_{a^2}$. In other words, we remove all actions that yield equal or worse treatment outcome but have higher cost when compared with another action in set A_{N-1}. For a given θ, let $G_{\beta_{N-1}}(\theta)$ denote the set of optimal solutions that satisfy $G_{\beta_{N-1}}(\theta) = \arg\max_{a \in A_{\beta_{N-1}}} V(a, \theta)$. Then, the actions that yield the largest $V(a, \theta)$ value are contained in $A_{\beta_{N-1}}$:

Lemma 3. $\arg\max_{a \in A_{N-1}} V(a, \theta) = G_{\beta_{N-1}}(\theta)$.

Proof. Proof of Lemma 3 For any fixed $\theta \in [0, 1]$, let a^1 and a^2 be two actions in set A_{N-1}. If $-C_{a^1} > -C_{a^2}$, and $H_{a^1} \leq H_{a^2}$, then $a^1 \notin A_{\beta_{N-1}}$, and we must have $V(a^1, \theta) < V(a^2, \theta)$. Thus, action a^1 will never appear in the optimal solution. ■ □

Recall that $a^*(\theta) = \arg\min_{a \in G_{\beta_{N-1}}(\theta)} H_a$. Given a set of effective actions $A_{\beta_{N-1}}$, without loss of generality, we sort the actions in $A_{\beta_{N-1}}$ according to their expected health from the smallest to the largest, and rename those actions from $1, \ldots, m$, where m is the cardinality of the set $A_{\beta_{N-1}}$. That is, $H_1 < H_2 < \ldots < H_m$. Thus, proving the first part of Claim 1 is equivalent to showing that $a^*(\theta)$ is non-decreasing as θ increases; this is because as $a^*(\theta)$ increases, $H_{a^*(\theta)}$ increases, and by the definition of the *effective action set*, $-C_{a^*(\theta)}$ also increases. To complete the proof of the first part of Property 2(b), it suffices to show the following propositions:

Proposition 3. *The optimal objective function value, $V(a^*(\theta), \theta)$, is non-decreasing with respect to θ.*

Proposition 4. *For a fixed β_{N-1}, given a sorted effective action set $A_{\beta_{N-1}}$ as indexed above, then the function $a^*(\theta)$ is non-decreasing with respect to θ.*

Proof. Proof of Proposition 3 Because $H_i > C_i$, for a fixed action a, the value of $V(a, \theta)$ is strictly increasing in θ. Thus, for every $\theta_i > \theta_j$, $\max_a V(a, \theta_i) = V(a^*(\theta_i), \theta_i) \geq V(a^*(\theta_j), \theta_i) \geq V(a^*(\theta_j), \theta_j) = \max_a V(a, \theta_j)$. ■ □

Before proving Proposition 4, we will first introduce another key concept that we will use throughout the rest of proof. Given two pairs of points (H_i, C_i) and (H_j, C_j) in the *effective action set*, with $i < j$, we define $\theta^*_{i,j} = \frac{C_i - C_j}{H_j - H_i + C_i - C_j}$, where $\theta^*_{i,j} \in (0, 1)$ is obtained by solving the equation $\theta H_i + (1 - \theta) C_i = \theta H_j + (1 - \theta) C_j$ for θ. Note that $\theta^*_{i,j}$ is well-defined when $i < j$ because by construction of the effective action set, $H_j > H_i$ and thus $-C_j > -C_i$. By the construction of $\theta^*_{i,j}$, we have:

Lemma 4. *when $\theta \in [0, \theta^*_{i,j})$, $V(i, \theta) > V(j, \theta)$. Similarly, when $\theta \in (\theta^*_{i,j}, 1]$, $V(i, \theta) < V(j, \theta)$.*

To complete the proof of Proposition 4, we need the following corollary implied by Lemma 4:

Corollary 1. *If we have a sequence of θ^*s satisfying $\theta^*_{x_1,x_2} \leq \theta^*_{x_2,x_3} \leq \ldots \leq \theta^*_{x_{m-1},x_m}$ with $x_k < x_{k+1} \forall k \in \{1, 2, \ldots, m-1\}$ where $x_k \in A_{\beta_{N-1}} \forall k$, then $a^*(\theta)$ is non-decreasing with respect to θ.*

Proof. Proof of Proposition 4 We will proceed by construction. Given a sorted effective action set $A_{\beta_{N-1}}$, we can find the solution to $V(a, \theta)$ by adding one action from $A_{\beta_{N-1}}$ at a time starting from the smallest indices. We show constructively that adding an action with higher health does not change the monotonicity of $a^*(\theta)$.

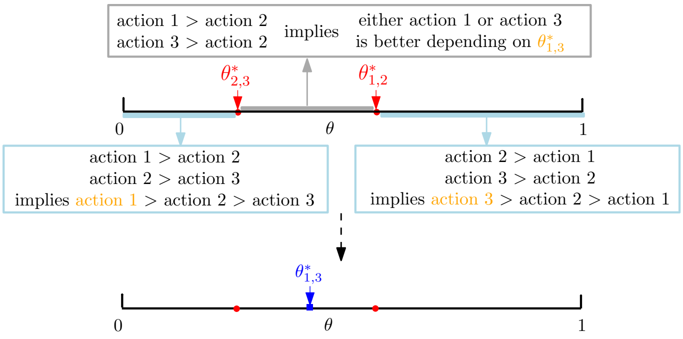


Figure 2.6: Proof of Proposition 4 step 2 case II.

Step 0: create an empty list L to keep track of the list of θ^* that are needed to select the optimal action given an input θ.

Step 1: add $\theta^*_{1,2}$ to L: by Lemma 4, we have monotonicity.

Step 2: add action 3 into the first two actions, $\{1, 2\}$, and calculate $\theta^*_{2,3}$.

Case I $\theta^*_{1,2} < \theta^*_{2,3}$: $a^*(\theta) = 1$ if $\theta \in [0, \theta^*_{1,2}]$, $a^*(\theta) = 2$ if $\theta \in (\theta^*_{1,2}, \theta^*_{2,3}]$, and $a^*(\theta) = 3$ if $\theta \in (\theta^*_{2,3}, 1]$. We add $\theta^*_{2,3}$ to L. Note that at $\theta^*_{i,j}$ with $i < j$ we choose action i by the definition of $a^*(\theta)$, that is, when there are actions with equivalent objective function under the same θ, we always pick the one with smaller expected health.

Case II $\theta^*_{2,3} < \theta^*_{1,2}$ (see Figure 2.6): when $\theta < \theta^*_{2,3}$, we have $V(2, \theta) > V(3, \theta)$, and $V(1, \theta) > V(2, \theta)$. This implies that $V(1, \theta) > V(2, \theta) > V(3, \theta)$. Thus, when $\theta \leq \theta^*_{2,3}$, $a^*(\theta) = 1$. Similarly, when $\theta > \theta^*_{1,2}$, we have $V(2, \theta) > V(1, \theta)$, and $V(3, \theta) > V(2, \theta)$. This implies that $V(3, \theta) > V(2, \theta) > V(1, \theta)$. Therefore, when $\theta \geq \theta^*_{1,2}$, $a^*(\theta) = 3$. Lastly, when $\theta^*_{2,3} < \theta < \theta^*_{1,2}$, we have $V(3, \theta) > V(2, \theta)$ and $V(1, \theta) > V(2, \theta)$. This implies that $a^*(\theta) = 1$ or 3. Thus, we calculate $\theta^*_{1,3}$. In this case, we add $\theta^*_{1,3}$ to L and remove $\theta^*_{1,2}$ from L (see Figure 2.6).

Case III $\theta^*_{1,2} = \theta^*_{2,3}$: $a^*(\theta) = 1$ if $\theta \in [0, \theta^*_{1,3}]$ and $a^*(\theta) = 2$ if $\theta \in (\theta^*_{1,3}, 1]$. We do not change L. The reasoning is the same as that in Case I.

In all cases, we have monotonicity.

Step i, $i \geq 3$: add action $i + 1$ to the previous actions and calculate $\theta^*_{i,i+1}$. Let $\theta^*_{j,i} \in L$ for some $j \in \{1, 2, ..., i-\}$. Note that this index j is in unique in L by construction.

Case I $\theta^*_{i,i+1} > \theta^*_{j,i}$: when $\theta \leq \theta^*_{i,i+1}$, $a^*(\theta)$ is monotone by previous steps; when $\theta > \theta^*_{i,i+1}$, $a^*(\theta) = i$. We add $\theta^*_{i,i+1}$ to L.

Case II $\theta^*_{i,i+1} < \theta^*_{j,i}$: then using the same logic as in step 2, we calculate $\theta^*_{j,i+1}$, add it to L and remove $\theta^*_{j,i}$ from L. We will now check whether $\theta^*_{j,i+1}$ violates the condition we defined in Corollary 1. We perform this procedure iteratively for all elements in L until no violation can be found. This procedure terminates in finite time because the number of elements in L is finite, and by construction, when it terminates, we obtain an L that satisfies Corollary 1.

Case III $\theta^*_{i,i+1} = \theta^*_{j,i}$: similar to Case III in step 2, we do not modify the set L, and $a^*(\theta)$ remains monotone according to the previous step.

■ □

Remark 1. *In the above we have shown that for every fixed belief state, the first part of Property 2(b) holds. Just to reiterate, this suffices to show Property 2(b) because the exact backward induction algorithms used in solving a POMDP relies on fixing the belief state at every searching point. (See § 2.5.3 for detailed discussion on one of the exact POMDP solution methods.) Note that since an MDP is a special case of POMDP with the belief state being standard unit vectors, this proof also applies to MDPs.*

So far, we have showed that as θ increases, the value of the objective function, $\theta H_{\beta_0}^{\pi_\theta^*} + (1-\theta) C_{\beta_0}^{\pi_\theta^*}$, the minimum optimal expected health, $H_{\beta_0}^{\pi_\theta^*}$, and the minimum absolute expected cost, $|C_{\beta_0}^{\pi_\theta^*}|$, are non-decreasing. The proof of Proposition 4 further implies that $|C_{\beta_0}^{\pi_\theta^*}|$ increases if and only if θ increases. To complete the proof of Property 2(b), it remains to show that as Γ_t increases, the optimal θ^* is non-decreasing. To observe this, we let γ_1 and γ_2 be two budget levels with $\gamma_1 > \gamma_2$. Because the set of all feasible policies under budget γ_2 is also feasible under budget γ_1, Equations (2.9)–(2.11) imply that the set of feasible θ under budget γ_2 is also feasible under budget γ_1. Let θ_1 and θ_2 be the optimal θ^* under budgets γ_1 and γ_2 respectively. Then, we must have $\gamma_1 \geq \gamma_2$ by Equation (2.9).

2.11.2 Proof of Property 2(c)

Proof. To show that $H_{\beta_0}^{\pi^*}(\Gamma_t)$ is non-decreasing in Γ_t, we proceed by contradiction. Let $\gamma_1 > \gamma_2$, and let the optimal policies of System (2) under $\Gamma = \gamma_1$ and $\Gamma = \gamma_2$ be π^1 and π^2, respectively. Let θ_1 and θ_2 be the optimal θ in Equation (2.9) under π^1 and π^2, respectively. Property 2(b) implies that since $\gamma_1 > \gamma_2$, $\theta_1 \geq \theta_2$. For the sake of contradiction, we assume $H_{\beta_0}^{\pi^1} < H_{\beta_0}^{\pi^2}$. Since π_2 is a feasible solution when $\Gamma_t = \gamma_1$, Equation (2.11) implies that

$$\theta_1 H_{\beta_0}^{\pi^1} + (1-\theta_1) C_{\beta_0}^{\pi^1} > \theta_1 H_{\beta_0}^{\pi^2} + (1-\theta_1) C_{\beta_0}^{\pi^2}. \tag{2.25}$$

Since we assume $H_{\beta_0}^{\pi^1} < H_{\beta_0}^{\pi^2}$, we must have $C_{\beta_0}^{\pi^1} \geq C_{\beta_0}^{\pi^2}$, that is $-C_{\beta_0}^{\pi^1} \leq -C_{\beta_0}^{\pi^2}$. In other words, policy π^1 is cheaper than policy π^2 but yields worse outcomes. This implies that π^1 is also a feasible solution when $\Gamma_t = \gamma_2$. Again, by Equation (2.11), when budget equals to γ_2, we have that

$$\theta_2 H_{\beta_0}^{\pi^1} + (1-\theta_2) C_{\beta_0}^{\pi^1} < \theta_2 H_{\beta_0}^{\pi^2} + (1-\theta_2) C_{\beta_0}^{\pi^2}. \tag{2.26}$$

Comparing Equations (2.25) and (2.26), we must have $\theta_2 > \theta_1$. Thus we have reached a contradiction.

■ □

2.12 Proof of Theorem 3

Theorem 3 (Correctness of our reformulation)**.** *Let $F(\Gamma_t)$ be piecewise linear, concave, and strictly increasing on $[l, u]$*[14] *as defined above, then $H^{\pi'}_{\beta_0} = H^{\pi^*}_{\beta_0}$ when $-C^{\pi'}_{\beta_0}$ lies on $F(\Gamma_t)$. That is, our reformulation, System II finds all solutions that lie on the efficiency frontier of the convex hull of the solutions of System I.*

Since the initial belief β_0 is fixed throughout this section, for a policy π, we will abbreviate $H^{\pi}_{\beta_0}$ with H_π and abbreviate $C^{\pi}_{\beta_0}$ with C_π. Because F is concave and strictly increasing on $[l, u]$, F is bijective. Thus, $F^{-1}(H_{\pi'_i})$ is well-defined and $F^{-1}(H_{\pi'_i}) = -C_{\pi'_i}$. Before we prove Theorem 3, we first show the following proposition:

Proposition 5. *Let $\theta^*_{\pi'_i,\pi'_j} = \frac{C_{\pi'_i}-C_{\pi'_j}}{H_{\pi'_j}-H_{\pi'_i}+C_{\pi'_i}-C_{\pi'_j}}$. If F is piecewise linear, concave, and strictly increasing, and $-C_{\pi'_j}$ is a non differentiable point on F, then $\theta^*_{\pi'_i,\pi'_j} < \theta^*_{\pi'_j,\pi'_k}$ $\forall \pi'_i, \pi'_j, \pi'_k$, s.t. $-C_{\pi'_i} < -C_{\pi'_j} < -C_{\pi'_k}$. In particular, there exists $\theta^*_{\pi'_j}$ s.t. $\theta^*_{\pi'_j} H_{\pi'_j} + (1-\theta^*_{\pi'_j})C_{\pi'_j} > \theta^*_{\pi'_j} H_\pi + (1-\theta^*_{\pi'_j})C_\pi$ for all policy $\pi \neq \pi'_j$.*

We observe that $\theta^*_{\pi'_i,\pi'_j}$ is simply an equivalence point of π'_i and π'_j in the objective function $\theta H_\pi + (1-\theta)C_\pi$ in System (2); that is, $\theta^*_{\pi'_i,\pi'_j} H_{\pi'_i} + (1-\theta^*_{\pi'_i,\pi'_j})C_{\pi'_i} = \theta^*_{\pi'_i,\pi'_j} H_{\pi'_j} + (1-\theta^*_{\pi'_i,\pi'_j})C_{\pi'_j}$. Furthermore, if $\theta > \theta^*_{\pi'_i,\pi'_j}$, then $\theta^*_{\pi'_i,\pi'_j} H_{\pi'_i} + (1-\theta^*_{\pi'_i,\pi'_j})C_{\pi'_i} < \theta^*_{\pi'_i,\pi'_j} H_{\pi'_j} + (1-\theta^*_{\pi'_i,\pi'_j})C_{\pi'_j}$, and if $\theta < \theta^*_{\pi'_i,\pi'_j}$, then $\theta^*_{\pi'_i,\pi'_j} H_{\pi'_i} + (1-\theta^*_{\pi'_i,\pi'_j})C_{\pi'_i} > \theta^*_{\pi'_i,\pi'_j} H_{\pi'_j} + (1-\theta^*_{\pi'_i,\pi'_j})C_{\pi'_j}$.

Proof. Proof of Proposition 5 Because $F^{-1}(H_{\pi'_i}) = -C_{\pi'_i}$, we rewrite $\theta^*_{\pi'_i,\pi'_j} = \frac{C_{\pi'_i}-C_{\pi'_j}}{F(-C_{\pi'_j})-F(-C_{\pi'_i})+C_{\pi'_i}-C_{\pi'_j}}$. We first notice that $\theta^*_{\pi'_i,\pi'_j}$ is well-defined for all π'_i, π'_j if $-C_{\pi'_i} > -C_{\pi'_j}$, and $\theta^*_{\pi'_i,\pi'_j} \in [0,1]$. Now,

$$\begin{aligned}
\theta^*_{\pi'_i,\pi'_j} < \theta^*_{\pi'_j,\pi'_k} &\iff \frac{C_{\pi'_i} - C_{\pi'_j}}{F(-C_{\pi'_j}) - F(-C_{\pi'_i}) + C_{\pi'_i} - C_{\pi'_j}} < \frac{C_{\pi'_j} - C_{\pi'_k}}{F(-C_{\pi'_k}) - F(-C_{\pi'_j}) + C_{\pi'_j} - C_{\pi'_k}} \\
&\iff \frac{F(-C_{\pi'_j}) - F(-C_{\pi'_i}) + C_{\pi'_i} - C_{\pi'_j}}{C_{\pi'_i} - C_{\pi'_j}} > \frac{F(-C_{\pi'_k}) - F(-C_{\pi'_j}) + C_{\pi'_j} - C_{\pi'_k}}{C_{\pi'_j} - C_{\pi'_k}} \\
&\iff \frac{F(-C_{\pi'_j}) - F(-C_{\pi'_i})}{C_{\pi'_i} - C_{\pi'_j}} > \frac{F(-C_{\pi'_k}) - F(-C_{\pi'_j})}{C_{\pi'_j} - C_{\pi'_k}}.
\end{aligned}$$

F is piecewise linear, concave, and strictly increasing, and $-C_{\pi'_j}$ is a non differentiable point; therefore, we have

$$\frac{F(-C_{\pi'_j}) - F(-C_{\pi'_i})}{C_{\pi'_i} - C_{\pi'_j}} \geq F'_-(-C_{\pi'_j}) > F'_+(-C_{\pi'_j})^{15} \geq \frac{F(-C_{\pi'_k}) - F(-C_{\pi'_j})}{C_{\pi'_j} - C_{\pi'_k}}.$$

Because this holds for all π_i, π_k s.t. $-C_{\pi'_i} < -C_{\pi'_j} < -C_{\pi'_k}$, without loss of generality, let $-C_{\pi'_i}$ be the non differentiable point right before $-C_{\pi'_j}$ and $-C_{\pi'_k}$ be the next non differentiable point after $-C_{\pi'_j}$;

[14] We pick l, u such that F is $-\infty$ on $[0, l)$ (i.e., infeasible) and constant on $[u, \infty)$

[15] $F'_-(x) := \lim_{t \longrightarrow 0^-} \frac{F(x+t)-F(x)}{t}$; $F'_+(x) := \lim_{t \longrightarrow 0^+} \frac{F(x+t)-F(x)}{t}$

then for any $\theta^*_{\pi'_j} \in (\theta^*_{\pi'_i,\pi'_j}, \theta^*_{\pi'_j,\pi'_k})$ satisfies $\theta^*_{\pi'_j} H_{\pi'_j} + (1 - \theta^*_{\pi'_j}) C_{\pi'_j} > \theta^*_{\pi'_j} H_\pi + (1 - \theta^*_{\pi'_j}) C_\pi$ for all policy $\pi \neq \pi'_j$. ■ □

Proof. Proof of Theorem 3 Let $\pi^*_{\theta^*}$ be an optimal policy in System II under budget $-C_{\pi^*_{\theta^*}}$ such that $-C_{\pi^*_{\theta^*}}$ corresponds to a non-differentiable point on $F(\Gamma_t)$ in System I on the X-axis. To prove Theorem 3, it suffices to show that $\pi^*_{\theta^*}$ is the optimal policy in System I when the budget equals $-C_{\pi^*_{\theta^*}}$. In other words, we want to show that Equations (2.9)-(2.11) find all non-differentiable points that lie on $F(\Gamma_t)$. Once we have established this, Equation (2.8) will ensure that we find all solutions that lie on the $F(\Gamma_t)$ and are differentiable. To complete the proof, we will proceed by contradiction.

First, the solution pair $(H_{\pi^*_{\theta^*}}, C_{\pi^*_{\theta^*}})$ is unique. For the sake of contradiction, assume that when the budget equals $-C_{\pi^*_{\theta^*}}$, there exists another policy $\tilde{\pi}$ s.t. $-C_{\tilde{\pi}} = -C_{\pi^*_{\theta^*}}$. If $H_{\tilde{\pi}} > H_{\pi^*_{\theta^*}}$, then we have $\theta^* H_{\tilde{\pi}} + (1 - \theta^*) C_{\tilde{\pi}} > \theta^* H_{\pi^*_{\theta^*}} + (1 - \theta^*) C_{\pi^*_{\theta^*}}$. This contradicts that $\pi^*_{\theta^*}$ is the optimal policy in System II. If $H_{\tilde{\pi}} < H_{\pi^*_{\theta^*}}$, then $\theta^* H_{\tilde{\pi}} + (1 - \theta^*) C_{\tilde{\pi}} < \theta^* H_{\pi^*_{\theta^*}} + (1 - \theta^*) C_{\pi^*_{\theta^*}}$ for all θ. Thus, policy $\tilde{\pi}$ is always dominated by policy π^* and cannot be an optimal solution.

Second, we want to show $H_{\pi'} = H_{\pi^*_{\theta^*}}$. To show $(H_{\pi'} \geq H_{\pi^*_{\theta^*}})$, we observe that all optimal policies in System II: $\{\pi^*_\theta \colon -C_{\pi^*_\theta} \leq \Gamma_t\}$, are feasible in System I. To show that $(H_{\pi'} \leq H_{\pi^*_{\theta^*}})$, we again proceed by contradiction. Note that all feasible policies in System I are still feasible in System II; however, since feasibility does not imply optimality in System II, we need to argue more carefully. Assume there exists an optimal solution of System I, π', that satisfies $H_{\pi'} > H_{\pi^*_{\theta^*}}$.

Case I $-C_{\pi'} \leq -C_{\pi^*_{\theta^*}}$: since π' is also feasible in System II, we have $\theta^* H_{\pi'} + (1 - \theta^*) C_{\pi'} > \theta^* H_{\pi^*_{\theta^*}} + (1 - \theta^*) C_{\pi^*_{\theta^*}}$. This contradicts that $\pi^*_{\theta^*}$ is the optimal policy.

Case II $\Gamma_t \geq -C_{\pi'} > -C_{\pi^*_{\theta^*}}$: Because $\pi^*_{\theta^*}$ is a optimal solution to $\arg\max_\pi \theta^* H_\pi + (1 - \theta^*) C_\pi$, we must have $\theta^* H_{\pi^*_{\theta^*}} + (1 - \theta^*) C_{\pi^*_{\theta^*}} \geq \theta^* H_{\pi'} + (1 - \theta^*) C_{\pi'}$. Because $H_{\pi'} > H_{\pi^*_{\theta^*}}$ and $-C_{\pi'} > -C_{\pi^*_{\theta^*}}$, Proposition 5 implies that there must exists[16] $\tilde{\theta} = \theta_{\pi^*_{\theta^*},\pi'} + \epsilon > \theta^*$,[17] with $\epsilon > 0$ s.t. $\tilde{\theta} H_{\pi^*_{\theta^*}} + (1 - \tilde{\theta}) C_{\pi^*_{\theta^*}} < \tilde{\theta} H_{\pi'} + (1 - \tilde{\theta}) C_{\pi'}$. This contradicts that $\theta^* = \max\{\theta : -C_{\pi^*_\theta} \leq \Gamma_t\}$. ■ □

2.13 Extension to Multiple Constraints

Kim et al. (2011) extends the definition of CPOMDP proposed in Isom et al. (2008) to multiple constraints, where a CPOMDP is defined as a POMDP with two additional components: 1) $c_k(s, a) < 0$ is the cost of type k incurred for executing action a in state s, and 2) $\gamma_k > 0$ is the upper

[16] the existence is implied by the second statement in Proposition 5.

[17] $\tilde{\theta} > \theta^*$ is implied by the first statement in Proposition 5.

bound on the absolute cumulative cost of type k. We can formally state these constraints as

$$\mathbb{E}_\pi \left[\sum_{t=1}^{N} -c_k(s,a) \right] \leq \gamma_t \quad \forall k.$$

Recall that when we have only one constraint, we reformulated our CPOMDP as follows:

$$\begin{aligned} \textbf{System II: } \pi^* &= \underset{\pi \in \Pi^*_{\theta^*}}{\arg\max} \left\{ -C^\pi_{\beta_0} \, : \, -C^\pi_{\beta_0} \leq \Gamma_t \right\} \\ \theta^* &= \max \left\{ \theta \, : \, -C^{\pi^*_\theta}_{\beta_0} \leq \Gamma_t, \theta \in [0,1] \right\} \\ \pi^*_\theta &= \underset{\pi \in \Pi^*_\theta}{\arg\min} \, H^\pi_{\beta_0} \\ \Pi^*_\theta &= \underset{\pi \in \Pi}{\arg\max} \left\{ \theta H^\pi_{\beta_0} + (1-\theta) C^\pi_{\beta_0} \right\}. \end{aligned}$$

We first illustrate how to incorporate two different types of cost constraints using this reformulation. The extension to the case where $k > 2$ is straightforward. To ease notation, we will omit the initial belief β_0. Let H^π denote the expected reward under policy π. Let C^π_1 and C^π_2 denote the expected cost of type 1 and type 2 under policy π, where C^π_1 and C^π_2 are always negative. We denote the upper bound on the absolute cumulative type 1 cost to be γ_1, and similarly γ_2 is the upper bound on the absolute cumulative type 2 cost. Furthermore, let $\theta, \eta \in [0,1]$ be two tunable parameters corresponding to the types 1 and 2 constraints, respectively. Then the idea is to solve two separate problems using System II, where the last steps in our budget reformulations are

$$\begin{aligned} \Pi^*_\theta &= \underset{\pi \in \Pi}{\arg\max} \left\{ (1-\theta) H^\pi + \theta C^\pi_1 \right\}, \\ \Pi^*_\eta &= \underset{\pi \in \Pi}{\arg\max} \left\{ (1-\eta) H^\pi + \eta C^\pi_2 \right\}, \end{aligned}$$

respectively. To solve our reformulation, we perform the following two steps:

Step 1 We solve the problem by pretending only constraint 1 exists (using System II) and find the corresponding optimal parameter $\widehat{\theta}^*$ and optimal solution $\widehat{\pi}^*_1$. Similarly, solve the problem using only constraint 2 and obtain $\widehat{\eta}^*$ and $\widehat{\pi}^*_2$. If one of those solutions also satisfies the other constraint, then we terminate and return that policy.

Step 2 If neither of the optimal solutions $\widehat{\pi}^*_1$ and $\widehat{\pi}^*_2$ are feasible, then we need to increase[18] the values of θ and η iteratively by a small amount until we find a policy that satisfies both constraints with the smallest possible values for θ and η.

[18] Note that in this new formulation, we have flipped the role of θ and $1-\theta$ in System II: we need to increase the value of θ here instead of decrease.

The correctness of this extension follows directly from the correctness of our algorithm. However, instead of performing a binary search over the value of θ when we only have one constraint, we now need to search over the grid of possible values of (θ, η) pairs in Step 2 above. As the number of constraints grows, the running time of our algorithm will grow exponentially with respective to the number of constraints in the problem, and the problem might become computationally intractable. A smarter search mechanism could be proposed in this scenario (e.g., using ideals from the EM algorithm, gradient descent, or bi-objective optimization with ϵ-constraints (Mavrotas 2009)), however this is beyond the scope of this problem.

2.14 Algorithms

Algorithms 3-5 are adopted from Cassandra et al. (1997). Let e_s be the standard basis vector that has 1 on the s^{th} entry and 0 everywhere else.

Algorithm 3 Lark's algorithm for purging a set of vectors

```
Lark Prune($A$)
$F \longleftarrow A$                                   ▷ dirty set
$Q \longleftarrow \emptyset$                                   ▷ clean set
for s in S do
    $\omega \longleftarrow \arg\max_{v \in F} e_s \cdot v$
    $Q \longleftarrow Q \cup \{\omega\}$
    $F \longleftarrow F \setminus \{\omega\}$
end for
while $F$ is not empty do
    for $v$ in $F$ do
        $x \longleftarrow$ Donimate(v,Q)
    end for
    if $x$ = donimtaed then
        $F \longleftarrow F \setminus \{v\}$
    else
        $\omega \longleftarrow \arg\max_{v \in F} x \cdot v$
        $Q \longleftarrow Q \cup \{\omega\}$
        $F \longleftarrow F \setminus \{\omega\}$
    end if
end whilereturn Q
```

Algorithm 4 An LP approach to finding an information state in a vector’s witness region

```
Dominate($\alpha, A$)
$L \longleftarrow LP$(variable : $x_1, ..., x_{|S|}, \delta$, objective : max $\delta$)
for $\alpha'$ in $A \setminus \{\alpha\}$ do
    add constraint $(L, x \cdot \alpha \geq \delta + x \cdot \alpha')$
    add constraint $(L, x \cdot \nmid = 1)$
end for
if Infeasible (L) then return dominated
else
    $(x, \delta) \longleftarrow$ SolveLP(L)
    if $\delta \leq 0$ then return $x$
    else return dominated
    end if
end if
```

Algorithm 5 Incremental pruning

```
Incremental Pruning($\mathscr{A}^{a,o_1}, ..., \mathscr{A}^{a,o_{|O|}}$)
$W \longleftarrow$ Lark Prune($\mathscr{A}^{a,o_1} \oplus \mathscr{A}^{a,o_3}$)
for i in $[3 : |O|]$ do
    $W \longleftarrow$ Lark Prune($W \oplus \mathscr{A}^{a,o_i}$)
end forreturn W
```

Algorithm 6 Group $\mathscr{A}_t^{a,o}$ into subsets according to d_v

```
regroup($\mathscr{A}_t^{a,o}$, $d_v$)
  $F \longleftarrow \mathscr{A}_t^{a,o}$    ▷ dirty set; each element in $F$ is a tuple of the form $(v, c)$
  $Q \longleftarrow$ array of $-1$ with length $|F|$    ▷ store the group index that each element in $F$ belongs to
  q_index $\longleftarrow 0$    ▷ initialize group index for $Q$
  for  i in range(len($F$)) do
      if  $Q[i] \neq -1$ then Continue    ▷ we have already assigned a group index to this element
      else $Q[i] \longleftarrow$ q_index
      end if
      for j in range(i, len(F)) do
          if $Q[j] \neq -1$ then Continue
          else if $\|Q[i].v - Q[j].v\|_1 < d_v$ then
              $Q[j] \longleftarrow$ q_index
          end if
      end for
      q_index + = 1    ▷ increment the group index
  end for
  return Q, q_index    ▷ return the group index of each element in F, and the total number of groups
```

2.15 Heuristic Algorithm for Solving Unconstrained POMDP

Algorithm 2, incorporating the removal of *α-vector*s that exceed the budget constraint, can be solved within several hours when the size of the problem is comparatively small. While a few heuristic methods have been proposed to speed up the run-time of finite-horizon POMDPs (Walraven and Spaan 2019), in this section we provide a heuristic algorithm to speed up the running time of our CPOMDP, Algorithm 2, when the complexity of the problem that we are trying to solve increases. This heuristic algorithm relies on reducing the size of the *α-vector* set, $\mathscr{A}_t^{a,o}$, described in Equation (2.20) in the pruning step. Note that: (i) the model described in § 2.4 of this paper can be solved exactly, and (ii) this heuristic algorithm could also be applied to unconstrained discrete-time finite-horizon POMDPs.

First, we observe that in our problem, when the size of the set $\mathscr{A}_t^{a,o}$ is large, the time needed to solve our CPOMDP increases, often because the absolute element-wise difference between some *α-vectors* is very small. Since the sets $\mathscr{A}_t^{a}$ and $\mathscr{A}_t$ are obtained by performing set operations on $\mathscr{A}_t^{a,o}$, reducing the size of $\mathscr{A}_t^{a,o}$ can also effectively reduce the size of $\mathscr{A}_t^{a}$ and $\mathscr{A}_t$. Thus, the idea of our heuristic algorithm is to remove the *α-vector* tuples that are “too close" to each other in the objective vector. We use the L1-norm as our distance metric, but it can be replaced with any p-norms for $p \geq 2$ in practice. In addition, since for a fixed vector x its p-norms, $\|x\|_p$, is monotonically decreasing with respect to p (Raıssouli and Jebril 2010), our theoretical guarantee below holds for all p-norms where $p \geq 2$.

Let m_v be a positive tunable parameter that control the threshold distance between two α-*vectors*. Algorithm 7 describes the heuristic that we use to reduce the size of $\mathscr{A}_t^{a,o}$, where $|O_p|$ is the number of all possible observations that we can have in a partially observable state. The function regroup($\mathscr{A}_t^{a,o}$, d_v) clusters the elements in $\mathscr{A}_t^{a,o}$ so that any two vectors within the same cluster have an L1 distance that is smaller than or equal to d_v (see § 2.14, Algorithm 6). Algorithm 7 can be used inside the *incremental pruning* (Algorithm 5 in § 2.14) to reduce the running time (from over 24 hours to 20 minutes on randomly generated instances).

Algorithm 7 Reduce the size of the $\mathscr{A}_t^{a,o}$ set

```
reduce($\mathscr{A}_t^{a,o}$, $m_v$, $|O_p|$)
$F \longleftarrow \mathscr{A}_t^{a,o}$          ▷ dirty set; each element in $F$ is a tuple of the form $(v, c)$
$d_v \longleftarrow m_v/2|O_p|$               ▷ initialize the cutoff distance for the objective function
$Q$, total_group ⟵ regroup($\mathscr{A}_t^{a,o}$, $d_v$)  ▷ $Q$ store the group index that each element in $F$ belongs to
$U \longleftarrow \emptyset$                                          ▷ clean set;
for $g$ in range(total_group) do
    $M$ ⟵ elements in F that are assigned to group $g$ in $Q$
    add a random element in $M$ to $U$
end for
return $U$
```

First, let V^* be the optimal value of the objective function of an unconstrained POMDP problem (Equation (2.11) of our reformulation) solved using the exact solution, and let $\widetilde{V}$ be the solution obtained by applying Algorithm 7 to Line (14) in Algorithm 1. Let l be the number of horizons that we apply Algorithm 7 to Algorithm 1 in Algorithm 2. Note that $l \leq N$, where N is the length of horizon. First, we show that the objective value that we obtain in Equation (2.11) by applying our heuristic algorithm (Algorithm 7) is at most $m_v \times l$ away from the optimal solution:

Theorem 4. $V^* - \widetilde{V} \leq m_v \times l$.

The proof of Theorem 4 can be found in § 2.15.1; this result holds for all unconstrained discrete-time, finite-horizon POMDPs.

2.15.1 Proof of Theorem 4

Theorem 4. $V^* - \widetilde{V} \leq m_v \times l$.

Proof. Proof of Theorem 4 Recall that V^* and $\widetilde{V}$ are the optimal values of an unconstrained POMDP problem solved using the exact solution and by applying *reduce*($\mathscr{A}_t^{a,o}$, m_v, 1), respectively. Let $\widehat{V}$ be the optimal value of the POMDP problem solved by applying our heuristic algorithm, Algorithm 7, to Line (18) of Algorithm 1, i.e., *reduce*($\mathscr{A}_t$, m_v, 1). We first show that $V^* - \widehat{V} \leq m_v \times l$ by induction, and then show that $V^* - \widetilde{V} \leq V^* - \widehat{V}$.

Base case $l = 1$: note that since the set $\mathscr{A}_N$ contains only a singleton according to Equation (2.15), without loss of generality, assume that we reduce the size of $\mathscr{A}_t$ at $t = N - 1$, that is, in the Second To the Last Step (STLS) of the backward induction. Let $\alpha^*_{N-1}(s)$ be the s^{th} entry of the α-vector that we removed from $\mathscr{A}_{N-1}$ but is used to compute V^* in the exact solution, and $\hat{\alpha}_{N-1}(s)$ be the value that is used to compute $\widehat{V}$. By construction, we have $\sum\limits_{s \in S} |\alpha^*_{N-1}(s) - \hat{\alpha}_{N-1}(s)| \leq m_v$. This implies that $|\alpha^*_{N-1}(s) - \hat{\alpha}_{N-1}(s)| \leq m_v \; \forall s$. Recall that by Equation (2.19), at time t equals to $N - 2$, any element inside $\mathscr{A}_{N-2}^{a_{N-2}, o_{N-1}}$ has the following form:

$$\tau(\alpha_{N-1}, a_{N-2}, o_{N-1})(s) = \frac{r(a_{N-2}, s)}{|O|} + \sum_{s' \in S} \alpha_{N-1}(s') w(o_{N-1}|s', a) \mathcal{P}(s'|s, a_{N-2}),$$

where $\frac{r(a_{N-2}, s)}{|O|}$ is a model specific constant, and α_{N-1} is some vector inside $\mathscr{A}_{N-1}$. Let $\alpha^{*,a}_{N-2}, \hat{\alpha}^a_{N-2} \in \mathscr{A}^a_{N-2}$ be the *α-vectors* that are used to compute V^* and $\widehat{V}$, respectively, in the set $\mathscr{A}^a_{N-2}$ at time $N - 2$. Let $\widehat{\mathscr{A}_{N-1}}$ denote the set of *α-vectors* that we obtained after applying Algorithm 7 to the set $\mathscr{A}_{N-1}$. Thus, we have $\alpha^{*,a}_{N-2} \notin \widehat{\mathscr{A}_{N-1}}$ and $\hat{\alpha}^a_{N-2} \in \widehat{\mathscr{A}_{N-1}}$. By Algorithm 1, $\alpha^{*,a}_{N-2}$ and $\hat{\alpha}^a_{N-2}$ are obtained by some summing over the set of all possible observations over $\tau(\alpha_{N-1}, a_{N-2}, o_{N-1})$, for some *α-vectors* $\alpha_{N-1} \in \mathcal{A}_{N-1}$. Abusing the notation a bit, let $\{\alpha^*_{x_i}\}_i$ and $\{\hat{\alpha}_{y_i}\}_i$ denote two sequences of *α-vectors* inside the set $\mathscr{A}_{N-1}$ and $\widehat{\mathscr{A}_{N-1}}$, respectively, where the index i corresponds to some observation o_i. (Note that $\alpha^*_{x_i}$ could equal $\alpha^*_{x_j}$ for $i \neq j$.) In particular, those two sequences of *α-vectors* were used to generated $\alpha^{*,a}_{N-2}$, and $\hat{\alpha}^a_{N-2}$ respectively, i.e., $\alpha^{*,a}_{N-2}(s) = \sum_{i=1}^{|O|} \tau(\alpha^*_{x_i}, a, o_i)(s)$ and $\hat{\alpha}^a_{N-2}(s) = \sum_{i=1}^{|O|} \tau(\hat{\alpha}_{y_i}, a, o_i)(s)$. Because $0 \leq w(o|s', a), \mathcal{P}(s'|s, a) \leq 1$, $\sum\limits_{s' \in S} \mathcal{P}(s'|s, a) = 1$, and $\sum\limits_{i=1}^{|O|} w(o_i|s', a) = 1$, we can bound the absolute difference between $\alpha^{*,a}_{N-2}(s)$ and $\hat{\alpha}^a_{N-2}(s)$ as follows:

$$\begin{aligned}
\left|\alpha^{*,a}_{N-2}(s) - \hat{\alpha}^a_{N-2}(s)\right| &= \left|\sum_{i=1}^{|O|} \sum_{s' \in S} \alpha^*_{x_i}(s') \mathcal{P}(s'|s, a) w(o_i|s', a) - \sum_{i=1}^{|O|} \sum_{s' \in S} \hat{\alpha}_{y_i}(s') \mathcal{P}(s'|s, a) w(o_i|s', a)\right| \\
&= \left|\sum_{s' \in S} \sum_{i=1}^{|O|} \left[\alpha^*_{x_i}(s') - \hat{\alpha}_{y_i}(s')\right] w(o_i|s', a) \mathcal{P}(s'|s, a)\right| \\
&\leq \sum_{s' \in S} \sum_{i=1}^{|O|} \left|\alpha^*_{x_i}(s') - \hat{\alpha}_{y_i}(s')\right| w(o_i|s', a) \mathcal{P}(s'|s, a) \\
&\leq \sum_{s' \in S} \sum_{i=1}^{|O|} m_v w(o_i|s', a) \mathcal{P}(s'|s, a) = m_v.
\end{aligned}$$

So far, we have shown that $|\alpha^{*,a}_{N-2}(s) - \hat{\alpha}^{a}_{N-2}(s)| \leq m_v \; \forall s$ given a fixed action a. Now let a^*_t and $\hat{a}_t$ be the optimal action that is chosen at time t in our exact solution and approximate solution, respectively.

Case I $a^*_t = \hat{a}_t \; \forall t \in \{N-3, ..., 1\}$: because $a^*_{N-1} = \hat{a}_{N-1}$, we have $|\alpha^*_{N-2} - \hat{\alpha}_{N-2}| \leq m_v$. Applying the same argument inductively, we can easily show that $\left|\alpha^{*,a}_t(s) - \hat{\alpha}^a_t(s)\right| \leq m_v$ for all t, and this implies that

$$V^* - \widehat{V} = \sum_{s \in S} \alpha^*_1(s)\beta_0(s) - \sum_{s \in S} \hat{\alpha}_1(s)\beta_0(s) \leq \sum_{s \in S} |\alpha^*_1(s) - \hat{\alpha}_1(s)|\beta_0(s) \leq m_v,$$

where $\beta_0(s)$ is the initial belief at state s with $\sum_{s \in S} \beta_0(s) = 1$.

Case II $a^*_t \neq \hat{a}_t$ for some $t \in \{N-3, ..., 1\}$: recall that during the reduction process, our heuristic algorithm only removed some elements in $\mathscr{A}_{N-1}$ (since $l = 1$). This implies that the set of feasible actions at every time step $t \in \{N-3, ..., 1\}$ remains the same. Thus, the action a^*_t is feasible at every time step t also for our heuristic algorithm. If the action a^*_t is indeed picked by our heuristic algorithm, then $\hat{a}_t = a^*_t$. By case I, we have our desired results, i.e., $V^* - \widehat{V} \leq m_v$. If another action is chosen instead in our heuristic algorithm, i.e., $\hat{a}_t \neq a^*_t$, then let $\overline{V}$ be the value of the policy where at each time t, we are forced to pick the same action as a^*_t. Because $\widehat{V}$ is the optimal solution in our heuristic algorithm, we must have $\widehat{V} \geq \overline{V}$. Using the proof from Case I, this implies that $V^* - \widehat{V} \leq m_v$.

Induction step: Assume that $V^* - \widehat{V} \leq m_v \times l$ is valid for $l = i$; show that the inequality holds when $l = i + 1$. From the base case, we observe that our worst case error bound holds constant if no additional reduction in the size of $\mathscr{A}_t$ for $t = 1, ..., N-2$ is performed. Thus without loss of generality, we can assume that the horizon of our POMDP problem has length $i + 1$. However, repeating the same argument as in the base case, we obtain that the worst case error increases by an additional factor of m_v.

To show that $V^* - \widetilde{V} \leq V^* - \widehat{V}$, we first claim that when the observation o is always observable, $|\mathscr{A}^{a,o}_t| = 1$ (we prove this below). When $|\mathscr{A}^{a,o}_t| = 1$, there is no need to reduce the size of $\mathscr{A}^{a,o}_t$. Thus, invoking *reduce*($\mathscr{A}^{a,o}_t, m_v, |O_p|$) when o is observable does not produce additional error. Consequently, the error on $\widetilde{V}$ is only induced by invoking *reduce*($\mathscr{A}^{a,o}_t, m_v, |O_p|$) for these sets where o is partially observable. More specifically, recall that the construction of algorithm *reduce*($\mathscr{A}^{a,o}_t, m_v, |O_p|$) implies that $\left|\tau\left(\alpha^*_{x_i}, a, o_i\right)(s) - \tau\left(\hat{\alpha}_{y_i}, a, o_i\right)(s)\right| \leq 2d_v = \frac{m_v}{O_p}$ for all states s. Thus, indeed in the base case above, at time t, we have

$$\begin{aligned}\left|\alpha^{*,a}_{N-2}(s) - \hat{\alpha}^a_{N-2}(s)\right| &= \left|\sum_{i=1}^{|O|} \left(\tau(\alpha^*_{x_i}, a, o_i)(s) - \tau(\hat{\alpha}_{y_i}, a, o_i)(s)\right)\right| = \left|\sum_{i=1}^{|O_p|} \left(\tau(\alpha^*_{x_i}, a, o_{p_i})(s) - \tau(\hat{\alpha}_{y_i}, a, o_{p_i})(s)\right)\right| \\ &\leq \sum_{i=1}^{|O_p|} \left|\tau\left(\alpha^*_{x_i}, a, o_{p_i}\right)(s) - \tau\left(\hat{\alpha}_{y_i}, a, o_{p_i}\right)(s)\right| \leq \frac{m_v}{|O_p|} \times |O_p| = m_v.\end{aligned}$$

Following the same induction above, we obtain the desired result.

To show that $|\mathscr{A}_t^{a,o}| = 1$ when o is observable, we notice $w(o|s', a)$ is either 1 or 0 for $\mathscr{A}_t^{a,o}$. In the case where $w(o|s', a) = 0$ when fixing a, o, then $\mathscr{A}_t^{a,o} = \{\vec{0}\}$; in the case where $w(o|s', a) = 1$, we are simply back to the MDP case where the optimal value of the state is unique. ■ □

2.16 Transition probability matrix

Recall the transition probability matrix from § 2.6.2:

$$\mathcal{P}_t^a = \begin{array}{c|cccccccc|} \mathit{from/to} & \text{Dx} & \text{NC} & \text{C1} & \text{C2} & \text{Re} & \text{OD} & \text{Dt} & \text{Abs} \\ \text{Dx} & x_{Dx} & nc_{Dx} & Z_1 & c^2_{Dx} & 0 & od_{Dx} & d_{Dx} & w \\ \text{NC} & 0 & nc_{NC} & Z_2 & 0 & 0 & 0 & d_{NC} & w \\ \text{C1} & 0 & nc_{C1} & Z_3 & c^2_{C1} & e_{C1} & od_{C1} & d_{C1} & w \\ \text{C2} & 0 & 0 & Z_4 & c^2_{C2} & e_{C2} & od_{C2} & d_{C2} & w \\ \text{Re} & x_{Re} & 0 & Z_5 & c^2_{Re} & e_{Re} & od_{Re} & d_{Re} & w \\ \text{OD} & 1 - d_{OD} & 0 & 0 & 0 & 0 & 0 & d_{OD} & 0 \\ \text{Dt} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \text{Abs} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array},$$

where $Z_1 = 1 - x_{Dx} - nc_{Dx} - c^2_{Dx} - od_{Dx} - d_{Dx} - w$, $Z_2 = 1 - nc_{NC} - d_{NC} - w$, $Z_3 = 1 - nc_{C1} - c^2_{C1} - e_{C1} - od_{C1} - d_{C1} - w$, $Z_4 = 1 - c^2_{C2} - e_{C2} - od_{C2} - d_{C2} - w$, $Z_5 = 1 - x_{Re} - c^2_{Re} - e_{Re} - od_{Re} - d_{Re} - w$.

Define: Nod: the number of overdoses that maps non-negative integer inputs to either 0 or 1.

Nr: the number of relapses that maps non-negative integer inputs to 1, 2, 3, 4, 5, or 6.

Tu: time since last use of drugs, measured in days.

TA: treatment adherence takes value 1(≥90%), 2(70%-90%), 3(<70%).

Let DL be the length of the detoxification program the patient chooses.

Variable	Action	Formula/Function of
x_{Dx}	action independent	$1 - \frac{1}{\text{DL}}$
nc_{Dx}	action independent	$(1 - x_{Dx} - od_{Dx} - d_{Dx} - w) \times 0.1$
c^2_{Dx}	action independent	$(1 - x_{Dx} - od_{Dx} - d_{Dx} - w) \times 0.7$

od_{Dx}	action independent	if $Nod > 0$: 0.524/2-year (Wines et al. 2007)
		if $Nod = 0$: 0.11/2-year (Wines et al. 2007)

d_{Dx}	action independent	if $Nr = 0$: $0.001/DL$ (Krebs et al. 2017)
		if $Nr > 0$: 0.001

w	action independent	0.02/year (Morral et al. (1997), O'Toole et al. (2006), TERMORSHUIZEN2005231)
nc_{NC}	action dependent	-
d_{NC}	action independent	0.005/263-day
nc_{C1}	action dependent	-
c^2_{C1}	action dependent	-
e_{C1}, e_{C2}	action dependent	one year relapse rate of 72% is obtained from (Chalana et al. 2016) the effect of treatment adherence is estimated from (Nosyk et al. 2009)
od_{C1}	action dependent	estimated from (Wines et al. 2007)
d_{C1}	action independent	0.008/263-day
c^2_{C2}	action dependent	-
od_{C2}	action dependent	estimated from (Wines et al. 2007)
d_{C2}	action independent	0.01/263-day
x_{Re}	NT	0.18/month
	M, B, IN, C	TA 1: 0.4/month;
	M, B, C	TA 2: 0.3/month; TA 3: 0.2/month
E	NT (Zarkin et al. 2005)	0.007/month
	M, B, IN, C (Zarkin et al. 2005)	TA 1: 0.018/month;
	M, B, C (Zarkin et al. 2005)	TA 2:0.016/month; TA 3: 0.0144/month
c^2_{Re}	NT (Zarkin et al. 2005)	0.004/month
	M, B, IN, C (Zarkin et al. 2005)	TA 1: 0.012/month;
	M, B, C (Zarkin et al. 2005)	TA 2:0.01/month; TA 3:0.0084/month
od_{Re}	action independent (Krebs et al. 2017)	$0.001 \times Nod + 0.009$
d_{Re}	action independent (Krebs et al. 2017)	0.021/263-day
d_{OD}	NT (Kelty and Hulse 2017), C	if $Tu \geq 29$ + DL: $0.158 + 0.01 \times Nod$
	M (Kelty and Hulse 2017)	if $Tu < 29$ + DL: $0.25 + 0.01 \times Nod$
	M (Kelty and Hulse 2017), IN	if $Tu \geq 29$ + DL: $0.11 + 0.01 \times Nod$
	B (Kelty and Hulse 2017), IN	if $Tu < 29$ + DL : $0.15 + 0.01 \times Nod$
	B (Kelty and Hulse 2017)	if $Tu \geq 29$ + DL : $0.09 + 0.01 \times Nod$

Table 2.6: The description of the parameters in the transition probability matrix. -: assumptions that we made to fill out the transition probability matrices.

2.17 Additional Figures

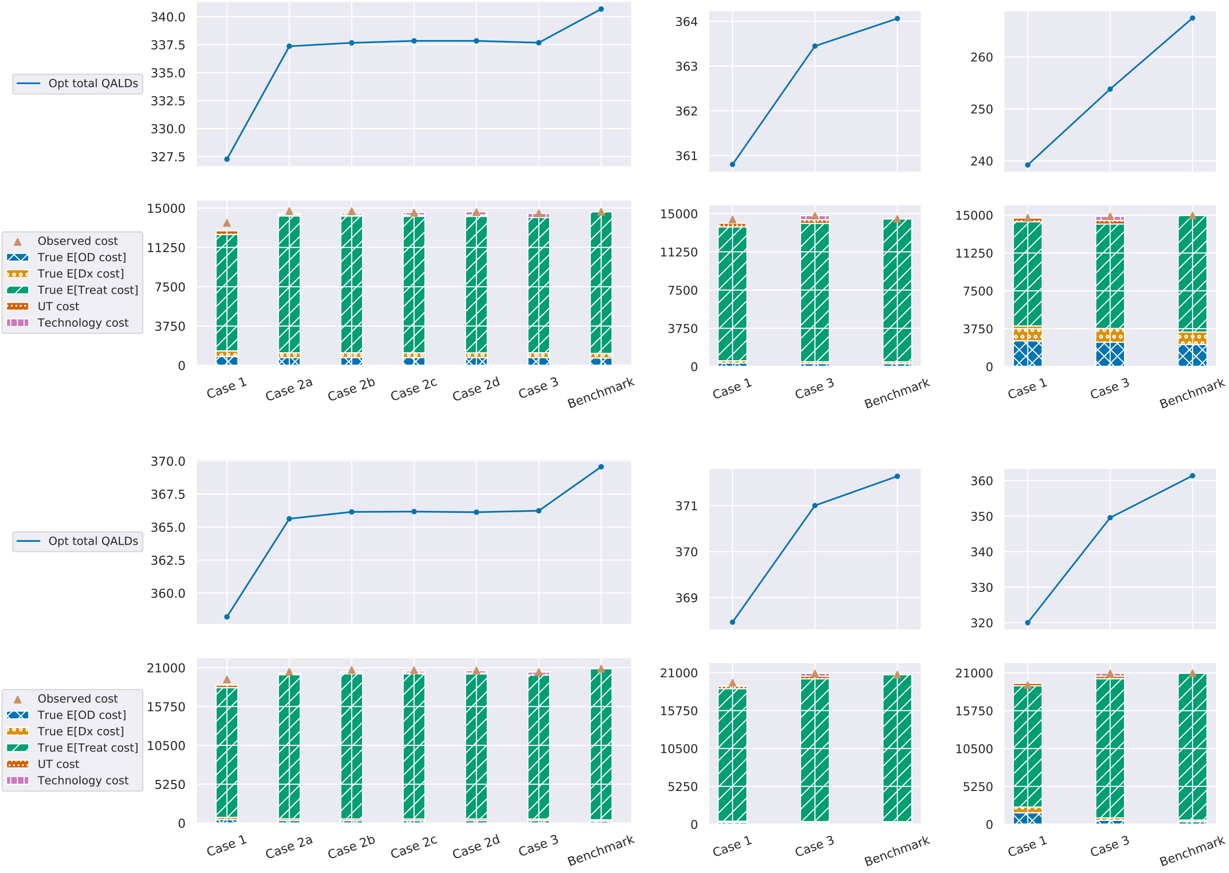


Figure 2.7: Additional Figures for Scenario 2 where the patient reacts to treatment B better than M but with a different magnitude than the average treatment dynamics. Top row: budget 15K. Bottom row: budget 21K. Left column: medium TA. Middle column: high TA. Right column: low TA. The optimal θ's, from top left to bottom right are approximately [0.294, 0.234, 0.201, 0.185, 0.162, 0.154, 0.158], [0.534, 0.421, 0.429], [0.126, 0.058, 0.058], [0.300, 0.272, 0.256, 0.239, 0.214, 0.202, 0.216], [0.551, 0.481, 0.483], and [0.146, 0.061, 0.061]; The expected QALDs received from treatments (excluding the terminal reward) are [284.91, 293.96, 294.88, 296.21, 296.32, 296.14, 298.78], [313.46, 315.26, 315.84], [229.57, 225.76, 237.29], [312.11, 318.27, 318.70, 318.72, 318.72, 318.80, 321.43], [320.14, 322.28, 322.86], and [297.28, 304.21, 314.31].

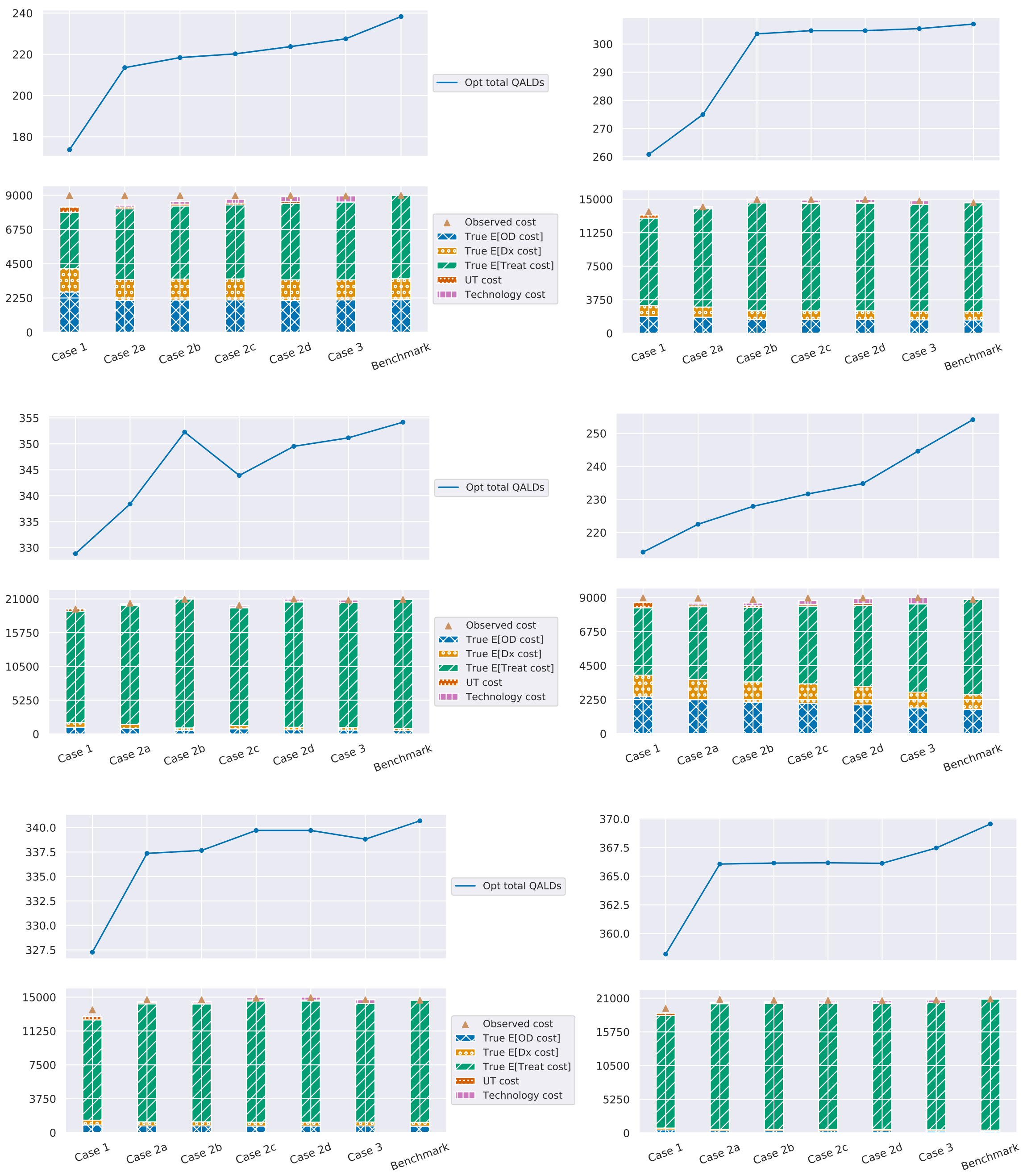


Figure 2.8: Reduced urine tests the frequency of urine tests in Cases 2a–2d is reduced to one third of the original frequency, and we remove urine tests in Case 3. In all subfigures, the patient has medium TA. Top row: Scenario 1. Top left: 9K budget. Top right: 15K budget. Middle left: 21K budget, Scenario 1. Middle right: 9K budget, Scenario 2. Bottom left: 15K budget, Scenario 2. Bottom right: 21K budget, Scenario 2. The optimal θ's, from top left to bottom right, are approximately [0.052, 0.065, 0.054, 0.045, 0.038, 0.037, 0.040], [0.294, 0.187, 0.139, 0.125, 0.110, 0.103, 0.103], [0.300, 0.222, 0.181, 0.165, 0.146, 0.138, 0.141], [0.052, 0.039, 0.036, 0.033, 0.026, 0.029, 0.029], [0.294, 0.234, 0.201, 0.185, 0.162, 0.158, 0.158], and [0.300, 0.274, 0.256, 0.239, 0.214, 0.208, 0.216].

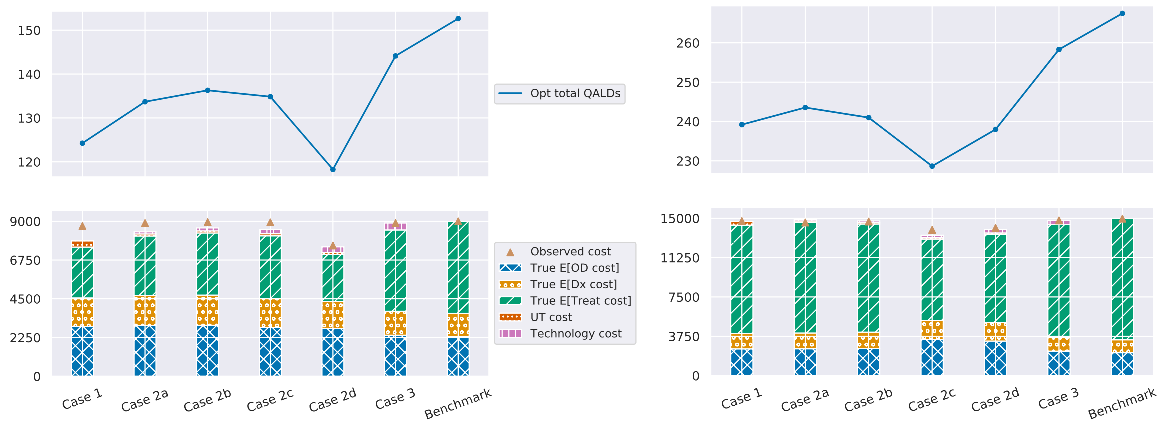


Figure 2.9: Additional figures for sensitivity analysis. Left: 9K budget, Scenario 1. Right: 15K budget, Scenario 2. the frequency of urine tests in Cases 2a–2d is reduced to one third of the original frequency, and we remove urine tests in Case 3. In both subfigures, the patient has low TA. Left: generate with a different matrix perturbation magnitude than that in Fig. 2.4(top right); right: generate with the same matrix perturbation as in Fig. 2.8. The optimal θ's, from left to right, are approximately [121.27, 130.31, 132.77, 129.19, 113.58, 132.38, 139.45] and [0.126, 0.081, 0.069, 0.065, 0.060, 0.058, 0.059]; the expected QALDs received from treatments (excluding the terminal reward) are [0.099, 0.065, 0.061, 0.061, 0.060, 0.056, 0.056] and [229.57, 232.98, 229.99, 212.17, 216.07, 229.55, 237.29].

2.18 Case Extension—A POMDP Formulation for Survey Devices

In this case, in addition to urine test results, we now have access to daily self-reports on cravings. However, a challenge arises because patients may not tell the truth all the time. In particular, a patient may provide true, noisy, or falsified information. We assume that monetary incentives are provided to encourage patients to respond to the surveys over 80% of the time (Serre et al. 2012). [19] We have the same setup for treatment dynamics, $\mathcal{P}_t^a$, and belief update, β_t, as in Case 1, but a different observation matrix, W. Notice that if a patient tells the truth all the time, the case will become equivalent to Case 4; if a patient provides only noisy information, this case is equivalent to Case 1. Consequently, we are interested only in the cases where patients provide falsified information or change their behaviors throughout the program.

To classify a patient's behavior, we first map the numbers 1 to 4 to the partially observable

[19] Any missing entries in the self-report of a patient can be regarded as either noisy or falsified information as needed in our case.

states by a function n: $n(\text{NC}) = 1, n(\text{C1}) = 2, n(\text{C2}) = 3$, and $n(\text{Re}) = 4$. We then find the worst health state (s_w) that a patient reported up to three days before the urine test, and compare it with the test result. If s_w matches the urine test, we say the patient is telling the truth. If $s_w = \text{Re}$, but we have a negative urine test result, we say the patient is providing noisy information. If $s_w \neq \text{Re}$, but the urine test result is positive, we say the patient is providing falsified information.[20] In particular, we define the distance $d(s_w, \text{Re}) = 4 - n(s_w)$. For example, if $s_w = \text{NC}$, then $d(\text{NC}, \text{Re}) = 4 - 1 = 3$. We also maintain an average distance and map the patient self-report state to the two closest states[21] accordingly: if the input label is NC and the average distance is 2.3, then we will map NC to C2 with a probability of 0.7 and to Re with a probability of 0.3.

We use the tuple (l_t, l_r, l_s) to keep track of the likelihood that a patient is telling the truth, providing noisy reports, or giving falsified information, where $l_t + l_r + l_s = 1$. After plugging in our estimated transition probability matrix in Case 1, $1 - nc_{ut+} > 0.99$ with nc_{ut+} defined in Case 1, so we assume that Re is fully observable in this case. The set of observations thus becomes patients' self-reports, and the observation matrix is $W = l_t \times W_T + l_r \times W_R + l_s \times W_S$, where $W_T = I_{8\times 8}$ is the observation matrix when a patient is completely truthful. W_R and W_S are the observation matrices when a patient is providing random information and strategic information, respectively, and they are defined as follows:

$$W_R = \begin{array}{c|cccccccc|} s/o & Dx & NC & C1 & C2 & Re & OD & Dt & Abs \\ \hline Dx & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ NC & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 \\ C1 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 \\ C2 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 \\ Re & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ OD & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ Dt & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ Abs & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}.$$

Let d be the average distance, $\lfloor d \rfloor$ be the integer part of d, and $\{d\} = d - \lfloor d \rfloor$ be the fractional part

[20] It could also be the case that the patient is providing noisy information. However, this scenario will be captured by parameter l_t below.

[21] If the average distance is an integer, we map it to only one label.

of d. Then

$$W_S(d|\lfloor d\rfloor = 1) = \begin{array}{c|cccccccc|}
s/o & Dx & NC & C1 & C2 & Re & OD & Dt & Abs \\
Dx & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
NC & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
C1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
C2 & 0 & \{d\} & 1-\{d\} & 0 & 0 & 0 & 0 & 0 \\
Re & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
OD & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
Dt & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
Abs & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}.$$

Similarly, we can enumerate the cases where $\lfloor d\rfloor = 0, 2, 3$:

$$W_S(d|\lfloor d\rfloor = 0) = \begin{array}{c|cccccccc|}
s/o & Dx & NC & C1 & C2 & Re & OD & Dt & Abs \\
Dx & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
NC & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
C1 & 0 & \{d\} & 1-\{d\} & 0 & 0 & 0 & 0 & 0 \\
C2 & 0 & 0 & \{d\} & 1-\{d\} & 0 & 0 & 0 & 0 \\
Re & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
OD & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
Dt & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
Abs & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array},$$

$$W_S(d|\lfloor d\rfloor = 2) = W_S(d|\lfloor d\rfloor = 3) = \begin{array}{c|cccccccc|}
s/o & Dx & NC & C1 & C2 & Re & OD & Dt & Abs \\
Dx & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
NC & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
C1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
C2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
Re & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
OD & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
Dt & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
Abs & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}.$$

To perform the belief update, we first calculate the observation matrix $W = l_t \times W_T + l_r \times W_R + l_s \times W_S$, and then follow the same procedures as in Case 1.

Chapter 3

Causal Inference with Selectively Deconfounded Data

3.1 Introduction

Say that we wished to determine whether sleep deprivation affects the likelihood of developing Alzheimer's disease. While we might easily observe that across a large population, sleep deprivation is highly *correlated* with Alzheimer's, this fact alone is not sufficient to establish whether sleep deprivation actually *causes* Alzheimer's. This is because the two variables might share a common cause (a *confounder*) accounting for the observed association, e.g., a genetic mutation that causes both sleep disorders and Alzheimer's.

The field of *causal inference* is concerned with precisely this type of problem. Generally, we wish to estimate the effect of assigning a given *treatment* (e.g., sleep deprivation) on a given *outcome* (e.g., Alzheimer's disease) in the presence of possible *confounders* (e.g., a genetic mutation).[1] One of the most fundamental problems in causal inference is to estimate the *average treatment effect* (ATE), i.e., the difference in the average outcomes that *would be* observed if everyone in the population did (versus did not) receive the treatment. The challenge of estimating the ATE in spite of unobserved confounding has been formalized mathematically (Pearl 1995, Rubin 1974), making clear that, in general, no amount of observational data is sufficient to identify the ATE. Specifically, there can (and typically do) exist multiple models consistent with the joint distribution over the observed variables that suggest different values of the ATE.

To bypass the challenge posed by confounding, research in causal inference has produced a variety of methods that tend to fall into two broad groups. The first group consists of performing

[1] Note here that in the technical sense, any variable upon which we could conceivably intervene and thereby influence the outcome can be called a *treatment*.

experiments on members of the population of interest, typically by assigning treatments randomly. The advantage of such *randomized controlled trials* (RCTs) is that they can eliminate unobserved confounding. However, while RCTs are often regarded as the gold standard for estimating causal effects, active experimentation is often infeasible, e.g., for ethical or financial reasons. The second group consists of leveraging domain knowledge to impose structural assumptions that render the parameter of interest identifiable. For example, if we assume that a sufficient set of confounders are observed (e.g., satisfying the backdoor criterion (Pearl 1995)), then under some mild statistical assumptions, the ATE can be estimated from observational data. Alternatively, even with unobserved confounding, other structural assumptions (e.g., valid instrumental variables or mediators) can render a causal effect estimable from observational data.

All of this is to say that the academic literature typically addresses the setting in which the confounders are either observed for *all* samples or for *none*. However, in many applications, it may be possible to revisit individuals represented in the dataset to retrospectively observe additional variables, such as the values of postulated confounders. For example, genetic information is stable over time, and thus when a genetic mutation is suspected to be a confounder, we might contact members of a study retrospectively to perform an additional test. Moreover, note that, as with this genetic example, even when the revelation of a variable is feasible it may nevertheless be expensive. In such settings, while it might be prohibitively expensive to reveal a confounder's value for every sample, we might still hope to reveal its value for a *selected subset* of our data. Consider the following examples:

1. **Drug Repositioning:** Here, drugs that have already been approved for a given disease are applied to treat a separate disease. Since these drugs are already proven to be safe, repositioning can be done in half the time and for a tenth of the cost when compared with the typical drug discovery process. To generate candidates for repositioning, correlations between drugs and diseases can easily be mined from health records. However, these relationships could plausibly be confounded by other drugs, which could both (a) affect the target disease and (b) trigger a side effect causing the candidate drug to be taken. While existing longitudinal studies aimed at assessing a drug's efficacy for a given disease A might have taken care to collect data on suspected confounders that might influence both treatment and disease A, when evaluating the drug as a candidate to treat disease B, we might need to reveal the value of other confounders. Moreover, due to the high costs of contacting patients and performing additional tests, such data collection efforts would be costly. Even for other known confounders, due to the digital divide of health records across disconnected systems, an expensive, manual effort, might be needed to collect the required data.

2. **Genetic Factors for Disease:** In the process of establishing the cause of different diseases,

genetic mutations are often implicated as potential confounders, such as in our initial motivating example for Alzheimer's. Due to the cost of DNA sequencing, it might only be possible to observe the genetic confounder for a subset of patients.

Succinctly, this paper addresses ATE estimation with *selectively deconfounded* data. We assume that we are given a large set of confounded data. At the outset only the treatment and outcomes are observed, but we have the ability to *deconfound* any of these samples, i.e., to reveal the sample's confounder. Naively, one could deconfound an arbitrary set of samples, and estimate the ATE with standard methods using only the subset of data that was deconfounded. However, this would discard a substantial amount of data. Thus motivated, we ask the following questions:

1. *What is the value of confounded data?* Specifically, how much can we improve ATE estimation by incorporating confounded data, relative to approaches that rely on deconfounded data alone?

2. *What is the additional value of selective deconfounding?* Specifically, how much can we further improve ATE estimation by intelligently selecting which confounded data to deconfound based upon the (observed) values of the treatments and outcomes?

To our knowledge, this is the first paper that focuses on the case where ample (cheaply-acquired) confounded data is available, and we may select only a limited number of samples to (at considerable cost) deconfound.[2]

3.1.1 Our Contributions

We address these questions for a standard confounding graph where the treatment and outcome are binary, and the confounder is categorical. More generally, we introduce a class of optimization problems that we dub *selective deconfounding*, where the values of the confounders are initially unobserved and can be subsequently revealed. Concretely, our ultimate goal is to find *sample-efficient selection policies* – policies for deciding which samples to deconfound in order to estimate the ATE most accurately. Our contributions are threefold:

1. The Value of Confounded Data (§ 3.3): We show (in Theorem 5) that a simple method for incorporating confounded data achieves a constant-factor improvement in ATE estimation error over using deconfounded data alone. Loosely, the reason for this is the following: in our non-parametric causal inference model, estimating the ATE boils down to estimating the parameters of

[2] Throughout this paper, we implicitly assume that the *confounded* data was sampled i.i.d. from the target population of interest (but that our policy for selecting data to deconfound need not be).

the data-generating distribution. The inclusion of (infinite) confounded data reduces the number of free parameters to be estimated, improving our estimates of the remaining parameters. Moreover, since the causal functional (the expression for the ATE) is non-linear in the parameters to be estimated, the error the estimated ATE can be much greater than the individual errors in parameter estimates. Thus, our improvements in parameter estimates yield greater benefits in estimating treatment effects. For binary confounders, our numerical results show that on average, over problem instances selected uniformly on the parameter simplex, our method achieves roughly a 2.5 factor reduction in ATE estimation error.

2. The Additional Value of Selective Deconfounding (§ 3.4): We show that selective deconfounding can in fact reduce ATE estimation error further, and we propose our own policy for doing so. We first establish (in Proposition 8) that there cannot exist an optimal policy for selective deconfounding, in the sense that no policy is universally optimal for all data-generating distributions. Thus, instead we compare our proposed policy against two benchmark policies (which we call "A" and "B" just for now), along different metrics. Assuming access to infinite confounded data, we find that:

1. With respect to the upper bounds on each policy's sample complexity that we show (Theorems 6 and 7), the guarantee for our proposed policy (a) entirely dominates benchmark policy A, and (b) is independent of (and therefore robust to) the observed distribution over treatment and outcome.

2. With respect to the (upper bound on) sample complexity of each policy under its *worst-case* data-generating distribution (Corollary 3.4), our proposed policy (a) dominates *both* benchmark policies, and (b) is independent (and again robust to) the entire data-generating distribution.

3. Among all estimators, we show (in Theorem 8) that our proposed policy requires no more than *twice* as many samples as benchmark policy B in the worst case, whereas benchmark policy B may require infinitely more samples as our policy.

We extend our work to the scenario where only a finite amount of confounded data is present, demonstrating that our qualitative insights continue to apply.

3. Experimental Evidence (§ 3.5): Our synthetic experiments suggest that our proposed policy dominates both benchmark policies when averaging over the unknown data-generating distributions. Moreover, our experiments also characterize those data-generating distributions most favorable/unfavorable for our policy. We validate our methods on COSMIC (Tate et al. 2019, Cosmic 2019), a real-world dataset containing cancer types, genetic mutations, and other patient features, showing that the practical benefits of our proposed sampling policy. We show that on average, to achieve an ATE estimation error of 0.006, our proposed selection policy reduces the number of deconfounded samples by a factor of up to 10 when compared with the two benchmark policies.

3.1.2 Related Work

Causal Inference Without Unobserved Confounders: Causal inference has been studied thoroughly under the ignorability assumption, i.e., no unobserved confounding (Neyman 1923, Rubin 1974, Holland 1986). Some approaches for estimating the ATE under ignorability include the backdoor adjustment (Pearl 1995, Huang and Valtorta 2006), using either outcome regression (Rubin 1974), inverse propensity score weighting (Rosenbaum and Rubin 1983, Hirano et al. 2003, McCaffrey et al. 2004), matching (Dehejia and Wahba 2002), or the use of instrumental variables when causal structural models are assumed (Sargan 1958, Angrist et al. 1996). Some related papers look to combine various sources of information, for instance from RCTs and observational data to estimate the ATE (Stuart et al. 2011, Hartman et al. 2015, Rosenman et al. 2018). Other papers leverage machine learning techniques, such as random forests, for estimating causal effects (Alaa and van der Schaar 2017, Wager and Athey 2018). Other techniques include using time-series data to estimate the ATE (Athey et al. 2016), and targeted learning (Van der Laan and Rose 2011). In the operations research and management science literature, several optimization problems have been proposed. Nikolaev et al. (2013) propose a new objective for matching such that the bias of the estimated treatment effect is minimized. Under resource constraints, Gupta et al. (2020) propose a new method for selecting individuals for treatment intervention to maximize the worst-case aggregate intervention effectiveness.

Under the ignorability assumption, missing data has also been thoroughly studied. Under the assumption that either the missing mechanism or the distribution of the complete data is correctly specified, doubly robust estimators (the estimators that remain consistent under this assumption) have been proposed when data is missing at random and when the missing probabilities are either known or can be parametrized (Robins et al. 1994, Hannah et al. 2010). Doubly-robust estimators have also been studied when outcome is missing not at random (Rotnitzky et al. 1998, Scharfstein et al. 1999), and been proposed for sequential (Robins 2000) and longitudinal (Bang and Robins 2005) missing data.

Causal Inference With Unobserved Confounders: Since an unaccounted for unobserved confounder can invalidate an estimate of the ATE, three lines of work attempt to address/remove

	Obs./exp. data	**Confounders**	**Guarantee**	**Active**
Robins et al. (1994)	Observational	Missing at random	Asymptotic	No
Kallus et al. (2018)	Both	Unconfounded exp. data	Asymptotic	No
This work	Observational	Selectively deconfounded	PAC	Yes

Table 3.1: Literature that are closely related to our paper

the ignorability assumption: one using observational data alone, another by combining confounded observational data with *experimental* (and thus unconfounded) data, and finally by conducting sensitivity analysis.

The first line includes papers using proxies (Miao et al. 2018) and mediators (Pearl 1995). Kuroki and Pearl (2014) identify graphical structures under which causal effect can be identified. Miao et al. (2018) propose to use two different types of proxies to recover causal effects with one unobserved confounder. Shi et al. (2018) extend the work by Miao et al. (2018) to multiple confounders. However, both methods require knowledge of proxy categories a priori and are not robust under misspecification of proxy categories. Louizos et al. (2017) use variational autoencoders to recover the causal effect under the model where when conditioned on the unobserved confounders, the proxies are independent of treatment and outcome. Pearl (1995) introduces the front-door adjustment, a procedure whereby the causal effect can be expressed as a functional that concerns only the (possibly confounded) treatment and outcome, and an (unconfounded) mediator that transmits the entire effect. This procedure was further improved by Tian and Pearl (2002) and Shpitser and Pearl (2006a,b).

The second line combines confounded observational and experimental data to estimate the ATE. Bareinboim and Pearl (2013) propose to combine observational and experimental data under distribution shift, learning the treatment effect from the experimental data and transporting it to the confounded observational data to obtain a bias-free estimator for the causal effect. Recently, Kallus et al. (2018) propose a two-step process to remove hidden confounding by incorporating experimental data, relaxing the assumption that the confounded data and experimental data have the same support region.

Finally, to better interpret the estimated ATE and to address the possibility of unobserved confounders, a third line of the work aims to test the robustness of the ATE estimate via sensitivity analysis. Note that this line of work does not address the existence of unobserved confounding, but rather examines how the estimated ATE would changed under one additional (hypothetical) confounder. These work include Cornfield et al. (1959), Rosenbaum (1987, 2002), Rosenbaum et al. (2010), Rosenbaum (2011, 2014), Shen et al. (2011), VanderWeele and Ding (2017), Zhao et al.

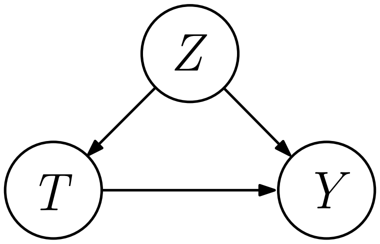


Figure 3.1: Causal graph with treatment T, outcome Y, and confounder Z

(2017), Miratrix et al. (2018). Fogarty and Small (2016) extend the analysis to multiple unobserved confounders by enforcing all unmeasured confounders to have the same impact on the treatment assignment probabilities.

Unlike most prior work, we (i) address confounded and deconfounded (but not experimental) observational data, (ii) perform finite sample analysis to quantify the relative benefit of additional confounded and deconfounded data towards improving our estimate of the average treatment effect, and (iii) investigate sample-efficient policies for selective deconfounding.

3.2 Model and Estimator

We begin by introducing the process through which (we assume) the data is generated and observed (Fig. 3.1). Let T and Y be binary random variables representing the *treatment* and *outcome*, respectively. The binary values can be taken, for example, as indicating whether a particular treatment has been applied (in the case of T) and whether the outcome was "successful" (in the case of Y). We assume the existence of a (potential) confounder, denoted Z, that can take up to k categorical values. Although there only exists a single confounder in our model, because the variable is categorical, the model subsumes scenarios with multiple categorical confounders. We will expand on this assumption later on in this section.

Let $\mathbb{P}_{Y,T,Z}$ denote the joint distribution of the random vector (Y, T, Z), the randomness corresponding to draws from a given population. We will use similar subscripts to denote conditional distributions (e.g., $\mathbb{P}_{Y|T,Z}$) and marginal distributions (e.g., $\mathbb{P}_Z$). *Confounded data* then consists of i.i.d. draws from $\mathbb{P}_{Y,T}$ (marginalized over the confounder Z), and *deconfounded data* consists of i.i.d. draws from $\mathbb{P}_{Y,T,Z}$. Thus, the confounded and deconfounded data are (y, t) and (y, t, z) tuples, respectively. We use the term *selective deconfounding* to mean selecting a confounded data point (y, t), and revealing the value of its confounder z – this corresponds to sampling z from $\mathbb{P}_{Z|Y,T}(\cdot|y, t)$.

Our goal is to estimate the *average treatment effect* (ATE). Now there are a number of (equivalent) formalizations of the ATE. One formalization follows the now-classical nomenclature established by Pearl (2000): let the so-called "do-conditional" expression $\mathbb{P}(Y = y|\text{do}(T = t))$ be defined as

$$\mathbb{P}(Y = y|\text{do}(T = t)) := \sum_{z\in[k]} \mathbb{P}_{Y|T,Z}(y|t, z)\mathbb{P}_Z(z).$$

Intuitively, $\mathbb{P}(Y = y|\text{do}(T = t))$ describes the probability of $\{Y = y\}$ when we "force" the treatment to take value t, regardless of what value it would have otherwise taken. For instance, in the Alzheimer's example, this corresponds to forcing an individual to sleep or not sleep.

The ATE then can be expressed, via the "back-door adjustment" (Pearl 1995), as

$$\begin{aligned} \text{ATE}(\mathbb{P}_{Y,T,Z}) &:= \mathbb{P}(Y = 1|\text{do}(T = 1)) - \mathbb{P}(Y = 1|\text{do}(T = 0)) && (3.1)\\ &= \sum_{z\in[k]} \left(\mathbb{P}_{Y|T,Z}(1|1, z) - \mathbb{P}_{Y|T,Z}(1|0, z)\right) \mathbb{P}_Z(z). && (3.2) \end{aligned}$$

In words, the ATE is the difference between the average outcome in the population were we to always administer the treatment, and the average outcome were we to never administer the treatment.

We make the following assumptions, all of which are by now standard in the causal inference literature:[3]

Assumption 1. *The following hold:*

1. ***Ignorability*** *[Rosenbaum and Rubin 1983]: $Y_t \perp\!\!\!\perp T|Z$, where Y_t is the outcome that Y would take if the treatment were t.*

2. ***Consistency*** *[Robins 1986]: The observed outcome for an individual i with treatment t, is the same as the one that we would observe if we were to assign treatment t to individual i.*

3. ***Positivity*** *[Cole and Hernán 2008]: $0 < \mathbb{P}_{T|Z}(1|z) < 1$ for all z such that $\mathbb{P}_Z(z) \neq 0$.*

Recall our motivation – we cannot calculate the ATE from confounded data alone.[4] Precisely, even given an *infinite* amount of confounded data (which corresponds to knowing $\mathbb{P}_{Y,T}$ exactly), there can exist multiple distributions $\mathbb{P}_{Y,T,Z}$ consistent with the confounded data, with different values for $\text{ATE}(\mathbb{P}_{Y,T,Z})$. So given confounded data alone, the only available options are to collect new deconfounded data, or selectively deconfound existing data.

[3] These assumptions are minimal: the ignorability assumption ensures the identifiability of the causal effect. The consistency rule is so ubiquitous in causal inference as to be axiomatic at this point (Pearl 2010). Finally, positivity ensures the estimability of the causal effect.

[4] We could if we intervene or make further assumptions on the structure of the causal graph, but these options are often unavailable, as in our motivating applications.

Aside: Model Generalizability: Our next step, in the coming subsection, will be to define the estimator for the ATE that we will study. But before that, we pause in this (optional) aside to address some of the components of our model that may appear to be restrictive at first glance. First, on the use of *discrete* variables: there do indeed exist applications and a rich literature, in which the variables are instead continuous. Now such variables can of course be quantized to fit our model, but more importantly, we note that the use of categorical (even binary) data is well-established in both theory (Bareinboim and Pearl 2013) and application (Knudson 2001, Rayner et al. 2016), and not merely a simplifying proxy for continuous data. Moreover, models with continuous variables, by and large, require some additional structural assumptions (e.g., linearity, an alternate parameterization, or smoothness). Our model requires no such assumption.

A second ostensible restriction is that we have a *single* confounder, whereas applications may have multiple confounders. Our model in fact subsumes these scenarios. In particular, absent additional distributional assumptions, our model captures multiple unobserved confounders by simple concatenation *without loss* (since we impose no limit on the number of classes). Now, one could make additional assumptions (indeed, a *high*-dimensional setting might necessitate such assumptions) that could render alternative algorithms applicable. However, there exist many applications where (a) the confounder is of moderate dimension; and (b) a practitioner would be dubious of any additional assumption (Bates et al. 2020).

Now concatenating multiple confounders into a single confounder does implicitly require that the set of confounders is either never observed or entirely observed, but this is also without loss so long as the costs of revealing the confounders are equal (e.g., this is the case in the genetic example). Intuitively, because we do not impose any independence assumption on the set of confounders, revealing all confounders offers maximal information on the joint distribution of the confounders. We formalize this statement in § 3.7.1.

Lastly, we have not considered pretreatment covariates. While for simplicity we focus only on the setting where the confounder can be retroactively observed, we show in § 3.7.2 that our model can be applied straightforwardly to handle additional pre-treatment covariates.

3.2.1 An ATE Estimator

We are ultimately concerned with the "value" of different sets of data: first to contrast combined (confounded and deconfounded) data against deconfounded data alone, and second to contrast different sets of selectively deconfounded data. Now the "value" of any data here is measured with respect to the accuracy with which that data can be used to estimate the ATE. It is worth emphasizing here that we are *not* directly concerned with the question of *how* data is used to estimate the ATE – the design of such ATE estimators is an active area of research (Robins et al.

1994, Robins 2000, Bang and Robins 2005), which we view as orthogonal to the questions we seek to answer. Our approach will be to fix a particular ATE estimator to be applied on all sets of data. We will now define this estimator, which is conceptually and algorithmically "simple," and corresponds to known, well-studied estimators (see § 3.2.2).

To ease notation slightly, let p^z_{yt} denote the probability of event $\{Y = y, T = t, Z = z\}$, let a_{yt} denote the probability of event $\{Y = y, T = t\}$, and let q^z_{yt} denote the conditional probability of event $\{Z = z|Y = y, T = t\}$:

$$p^z_{yt} := \mathbb{P}_{Y,T,Z}(y, t, z), \quad a_{yt} := \mathbb{P}_{Y,T}(y, t), \quad q^z_{yt} := \mathbb{P}_{Z|Y,T}(z|y, t).$$

We will further compact the notation by letting $\mathbf{p}$ denote the vector of all $4k$ (recall that Y and T are binary-valued, and Z takes one of k categorical values) values p^z_{yt}, in arbitrary order, and similarly for $\mathbf{a}$ and $\mathbf{q}$. As a sanity check, the ATE can be computed entirely from $\mathbf{p}$, but the reason to call out the decomposition $\mathbf{a}$ and $\mathbf{q}$ (it is a "decomposition" in the sense that $p^z_{yt} = a_{yt}q^z_{yt}$) is that it will highlight the different information contained in confounded data ($\mathbf{a}$) versus deconfounded data ($\mathbf{q}$). Now on to our estimator, which differs slightly based on the type of data being used:

1. **Deconfounded data alone**: Given only deconfounded data, we first obtain empirical estimates for each p^z_{yt} using the *maximum likelihood estimator* (MLE), which is simply the empirical frequency, and we denote the corresponding estimates $\hat{p}^z_{yt}$. We then obtain our estimated average treatment effect, $\widehat{\text{ATE}}$, by plugging $\hat{p}^z_{yt}$ directly into the definition of ATE in Eq. (3.2). Specifically, let $\text{ATE}(\mathbf{p}) : [0, 1]^{4k} \longrightarrow [-1, 1]$ denote the value of the ATE under the distribution $\mathbf{p}$:

$$\text{ATE}(\mathbf{p}) = \sum_z \left(\frac{p^z_{11}}{\sum\limits_y p^z_{y1}} - \frac{p^z_{10}}{\sum\limits_y p^z_{y0}} \right) \left(\sum_{y,t} p^z_{yt} \right) \tag{3.3}$$

(this is just a re-writing of Eq. (3.2)). Then our estimator under deconfounded data alone:

$$\widehat{\text{ATE}} = \text{ATE}(\hat{\mathbf{p}}). \tag{3.4}$$

While this estimator is quite simple, in § 3.2.2 we show that absent additional causal structural assumptions, it corresponds to the well-studied doubly-robust estimation method.

2. **Incorporating confounded data**: To assess the value of confounded data, recall we have decomposed $\mathbb{P}_{Y,T,Z}$ into two components: (i) the confounded distribution $\mathbb{P}_{Y,T}$, or $\mathbf{a}$; and (ii) the conditional distribution $\mathbb{P}_{Z|Y,T}$, or $\mathbf{q}$. The process of *deconfounding* reveals the value of Z for one (initially confounded) sample, and so we gain no additional information about the joint distribution $\mathbb{P}_{Y,T}$. Thus, $\mathbf{a}$ is estimated entirely using the confounded data, and the deconfounded data can then be used exclusively to estimate the conditional distribution $\mathbf{q}$.

(a) **Infinite samples of confounded data**: Much of our analysis concerns the case where the amount of confounded data is effectively infinite, so that the confounded distribution $\mathbf{a}$ is known exactly. Analogous to Eq. (3.3), we let $\text{ATE}(\mathbf{a}, \mathbf{q}) : [0,1]^4 \times [0,1]^k \longrightarrow [-1,1]$ denote the value of the ATE under the distribution decomposed as $\mathbf{a}$ and $\mathbf{q}$:

$$\text{ATE}(\mathbf{a}, \mathbf{q}) = \sum_z \left(\frac{a_{11} q_{11}^z}{\sum_y a_{y1} q_{y1}^z} - \frac{a_{10} q_{10}^z}{\sum_y a_{y0} q_{y0}^z} \right) \left(\sum_{y,t} a_{yt} q_{yt}^z \right). \tag{3.5}$$

We estimate each value q_{yt}^z from deconfounded data using the MLE estimator, and denote the estimates as $\hat{q}_{yt}^z$. We then calculate our estimated ATE by plugging the a_{yt}'s and $\hat{q}_{yt}^z$'s into Eq. (3.5):

$$\widehat{\text{ATE}} = \text{ATE}(\mathbf{a}, \hat{\mathbf{q}}). \tag{3.6}$$

(b) **Finite samples of confounded data**: Finally, in the case where we have a finite number of samples of confounded data, we use the confounded data to produce estimates $\hat{a}_{yt}$, via the MLE, and calculate our estimate $\widehat{\text{ATE}}$ by plugging the $\hat{a}_{yt}$'s and $\hat{q}_{yt}^z$'s into Eq. (3.5):

$$\widehat{\text{ATE}} = \text{ATE}(\hat{\mathbf{a}}, \hat{\mathbf{q}}). \tag{3.7}$$

3.2.2 Aside: Our Estimator as the Doubly Robust Estimator

In this (optional) subsection, we show that when incorporating *deconfounded data only*, without making any additional assumptions about the causal structure, our estimator, Eq. (3.4), is the same as the one obtained by applying the well-established doubly-robust estimation method, i.e. there is no benefit of doubly robustness under our problem setup. We do so by showing our estimator is the same 1) as the one obtained through a generic outcome regression model, and 2) as the one obtained via the *inverse-propensity score weighting* (IPW) method. In addition, we note that under our current problem setup, the IPW estimator *always* yield the same estimators as the ones in our paper, Eqs. (3.4) - (3.7). Finally, in § 3.8, we show that with infinite confounded data, a straightforward extension of the outcome regression model leads to an optimization problem that is not well-defined. This motivates us to provide estimator-independent theoretical guarantees in § 3.3.

Our Estimator as the Outcome Regression Estimator First, under deconfounded data only, we observe only the tuples (y, t, z) (and we do not know the confounded distribution $\mathbb{P}_{Y,T}$). Without making any additional causal structural assumptions, we can only estimate the *conditional average* *treatment effect* (CATE) – the treatment effect when conditioned on the value of the confounder Z – by one-hot encoding the (t, z) tuples. We can then obtain the ATE by reweighting the CATE by $\mathbb{P}(Z)$ (which can be estimated separately). We formally state the outcome regression process below.

we first note that the ATE when conditioned on $Z = z$ equals to $\mathbb{P}_{Y|T,Z}(1|1, z) - \mathbb{P}_{Y|T,Z}(1|0, z)$. Let u_{tz} be a binary random variable that takes the value 1 if $T = t$ and $Z = z$, and 0 otherwise. Let y_i^{tz} to denote the values of Y when $T = t$ and $Z = z$, where $i = 1, ..., N_{tz}$ and N_{tz} is the number of samples where $T = t$ and $Z = z$. Then, using the random variables u_{tz}'s, we can estimate the value of $\mathbb{P}_{Y|T,Z}(1|1, z)$ by running the following logistic regression: $\mathbb{P}_{Y|T,Z}(1|1, z) = \sigma(w_{tz}u_{tz} + b_{tz})$, where w_{tz} and b_{tz} are the weight and bias respectively, and σ is the logistic function. Then, the MLE for $\sigma(w_{tz}u_{tz} + b_{tz})$ can be obtained through solving the following optimization problem:

$$\arg\max \sum_{i=1}^{N_{tz}} y_i^{tz} \log(\sigma(w_{tz}u_{tz} + b_{tz}) + (1 - y_i^{tz}) \log(1 - \sigma(w_{tz}u_{tz} + b_{tz})),$$

and we obtain the MLE estimate $\tilde{Y}^{tz} = \frac{\sum_{i=1}^{N_{tz}} y_i^{tz}}{N_{tz}}$. Recall that $\hat{p}_{yt}^z$ is our estimated $\mathbb{P}_{Y,T,Z}(y, t, z)$ (using the MLE estimator), which can be calculated by dividing the number of samples where $Y = y, T = t, Z = z$ by N_{tz}. Let N be the total number of deconfounded samples. Then,

$$\frac{\sum_{i=1}^{N_{tz}} y_i^{tz}}{N_{tz}} = \frac{N\hat{p}_{1t}^z}{N\hat{p}_{0t}^z + N\hat{p}_{1t}^z} = \frac{\hat{p}_{1t}^z}{\hat{p}_{0t}^z + \hat{p}_{1t}^z}.$$

Thus, the estimated CATE under the above logistic regression can be expressed as

$$\mathbb{P}_{Y|T,Z}(1|1, z) - \mathbb{P}_{Y|T,Z}(1|0, z) = \frac{\sum_{i=1}^{N_{1z}} y_i^{1z}}{N_{1z}} - \frac{\sum_{i=1}^{N_{0z}} y_i^{0z}}{N_{0z}} = \frac{\hat{p}_{11}^z}{\hat{p}_{01}^z + \hat{p}_{11}^z} - \frac{\hat{p}_{10}^z}{\hat{p}_{00}^z + \hat{p}_{10}^z}.$$

One could verify that the above expression is indeed the same as the estimated CATE that we would have obtained using Eq. (3.4). Finally, to obtain the ATE, we can estimate $\mathbb{P}_Z$ separately using deconfounded data alone via the MLE and obtain that $\mathbb{P}_Z(z) = \sum_{y,t} \hat{p}_{yt}^z$. Together, we conclude that outcome regression yields the same estimator as in Eq. (3.4).

Our Estimator as the IPW Estimator Next, we show that under deconfounded data alone, the IPW estimator yields the same estimator as in Eq. (3.4).

Let $\hat{\tau}$ denote the estimated ATE. We define the propensity score $e(z) := \mathbb{P}(T = 1|Z = z)$, and let $\hat{e}(z)$ be the estimated propensity score from the data. Let t_i, y_i, and z_i denote the value of the treatment, outcome, and confounder for the i-th sample respectively. First, we recall that the IPW estimator takes the following form:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \frac{t_i y_i}{\hat{e}(z_i)} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - t_i) y_i}{1 - \hat{e}(z_i)}. \tag{3.8}$$

Note that in our problem, Z takes a finite number of discrete values. Thus, using the notation above we can further decompose Eq. (3.8) as follows:

$$\begin{aligned}
\hat{\tau} &= \frac{1}{N}\sum_z \left(\sum_{i=1}^{N_{1z}} \frac{y_i^{1z}}{\hat{e}(z)}\right) - \frac{1}{N}\sum_z \left(\sum_{i=1}^{N_{0z}} \frac{y_i^{0z}}{1-\hat{e}(z)}\right) && (3.9)\\
&= \sum_z \left(\frac{1}{\hat{e}(z)}\frac{\sum_{i=1}^{N_{1z}} y_i^{1z}}{N}\right) - \sum_z \left(\frac{1}{1-\hat{e}(z)}\frac{\sum_{i=1}^{N_{0z}} y_i^{0z}}{N}\right) && (3.10)\\
&= \sum_z \left(\frac{1}{\hat{e}(z)}\frac{N\hat{p}_{11}^z}{N}\right) - \sum_z \left(\frac{1}{1-\hat{e}(z)}\frac{N\hat{p}_{10}^z}{N}\right) = \sum_z \left(\frac{1}{\hat{e}(z)}\hat{p}_{11}^z\right) - \sum_z \left(\frac{1}{1-\hat{e}(z)}\hat{p}_{10}^z\right), && (3.11)
\end{aligned}$$

where the second to the last equality is due to the fact that $\sum_{i=1}^{N_{tz}} y_i^{tz} = N\hat{p}_{1t}^z$. Now, to estimate $\hat{e}(z) = \hat{\mathbb{P}}_{T|Z}(1|z)$, we note that without making any additional causal structural assumptions, $\hat{e}(z)$ can be expressed using our estimators as follows: $\hat{e}(z) = \frac{\sum_y \hat{p}_{y1}^z}{\sum_{yt}\hat{p}_{yt}^z}$, and $1 - \hat{e}(z) = \hat{\mathbb{P}}_{T|Z}(0|z) = \frac{\sum_y \hat{p}_{y0}^z}{\sum_{yt}\hat{p}_{yt}^z}$. Plugging in these values back into Eq. (3.11), we recover the Eq. (3.4) exactly. Finally, we observe that the IPW estimator can be formally stated as

$$\sum_z \frac{\mathbb{P}_{Y,T,Z}(1,1,z)}{\mathbb{P}_{T|Z}(1,z)} - \sum_z \frac{\mathbb{P}_{Y,T,Z}(1,0,z)}{\mathbb{P}_{T|Z}(0,z)}. \tag{3.12}$$

Thus, under our problem setup, without making any additional assumption, Eq. (3.12) *always* yields the same estimate as Pearl's backdoor adjustment formula Eq. (3.2).

3.3 The Value of Confounded Data

We are now prepared to answer our first question: *how much can we improve ATE estimation by incorporating confounded data*. To that end, we first analyze and compare the *sample complexity* of the MLE estimator described in § 3.2 for deconfounded data alone and for augmented with an infinite amount of confounded data, while holding everything else the same. Throughout the paper, we measure the sample complexity of an estimator $\widehat{\text{ATE}}$ as the number of samples required for $\widehat{\text{ATE}}$ to be (ϵ,δ)-close to the true ATE, where the notion of "(ϵ,δ)-close" is defined as follows:

Definition 2. *An estimator $\widehat{\text{ATE}}$ is said to be (ϵ,δ)-close to the true ATE if it satisfies*

$$\mathbb{P}(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon) < \delta.$$

Deconfounded Data Alone: We begin with the *baseline* approach of using only deconfounded data that has been sampled according to the deconfounded distribution, $\mathbb{P}_{Y,T,Z}$. The following theorem identifies a quantity of samples m_{base} which is sufficient to estimate the ATE to within a desired level of accuracy under the estimation process described above.

Theorem 5. *Using deconfounded data alone, the estimator* $\mathrm{ATE}(\hat{\boldsymbol{p}})$ *as defined in Equation* (3.4) *is* (ϵ, δ)*-close if the number of deconfounded samples is at least*

$$m_{\text{base}} := C \max_{t,z} \left(\sum_y p^z_{yt} \right)^{-2} = C \max_{t,z} \frac{1}{\mathbb{P}_{T,Z}(t,z)^2},$$

where $C := 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$.

The proof of Theorem 5 (in § 3.9.1) relies on an additive decomposition of the estimation error on ATE in terms of the estimation error on the p^z_{yt}'s, along with concentration via Hoeffding's inequality.

Incorporating Infinite Confounded Data: Now consider the setup in which we have deconfounded data, along with an *infinite* amount of confounded data, i.e., the marginal distribution $\mathbb{P}_{Y,T}$ is known exactly. Analogous to Theorem 5, Theorem 6 identifies a sufficient number of deconfounded samples m_{nsp}.[5]

Theorem 6. *Incorporating (infinite) confounded data, the estimator* $\mathrm{ATE}(\boldsymbol{a}, \hat{\boldsymbol{q}})$ *is* (ϵ, δ)*-close if the number of deconfounded samples is at least*

$$m_{\text{nsp}} := C \max_{t,z} \frac{\sum_y a_{yt}}{\left(\sum_y a_{yt} q^z_{yt}\right)^2} = C \max_{t,z} \frac{\mathbb{P}_T(t)}{\mathbb{P}_{T,Z}(t,z)^2}, \tag{3.13}$$

where $C := 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$.

The proof of Theorem 6 is included in § 3.9.5. A few observations can be made from comparing Theorems 5 and 6:

1. m_{nsp} is less than m_{base} for *any* underlying distribution $\mathbb{P}_{Y,T,Z}$, highlighting the value of confounded data. In fact, the ratio $m_{\text{base}}/m_{\text{nsp}}$, in addition to being strictly greater than 1, can be arbitrarily large.

2. With respect to the accuracy parameters ϵ and δ, both m_{base} and m_{nsp} scale as $\Omega(\epsilon^{-2} \ln(\delta^{-1}))$, irrespective of the underlying distribution $\mathbb{P}_{Y,T,Z}$. Now it might be that a better scaling is achievable, either with a "smarter" estimator or a tighter analysis of our estimator, but the following minimax lower bound shows that this is *not* the case:

   Proposition 6. *(Lower Bound with respect to* ϵ *and* δ*) Fix any confounded distribution and assume that infinite confounded data is given (or equivalently,* $\mathbb{P}_{Y,T}$ *is known). For any* ATE

[5] The subscript "nsp" stands for *natural selection policy.* The motivation for this name will be explained in the next section.

estimator, there exists an underlying distribution $\mathbb{P}_{Y,T,Z}$*, whose confounded distribution is* $\mathbb{P}_{Y,T}$*, for which the number of deconfounded samples*[6] *required for the estimator to be* (ϵ, δ)*-close is at least* $\Omega(\epsilon^{-2}\log(\delta^{-1}))$*.*

The proof of Proposition 6 (§ 3.9.3) proceeds by construction. Proposition 6 states not only that Theorems 5 and 6 are tight with respect to ϵ and δ, but moreover the $\Omega(\epsilon^{-2}\log(\delta^{-1}))$ sample-complexity is necessary even for *selectively* deconfounded data, as we will study in the next section.

3. Even fixing k, ϵ, and δ, both m_{base} and m_{nsp} can be arbitrarily large for certain underlying distributions $\mathbb{P}_{Y,T,Z}$. In fact, even fixing *any* confounded distribution $\mathbb{P}_{Y,T}$, both m_{base} and m_{nsp} can *still* be arbitrarily large for distributions $\mathbb{P}_{Y,T,Z}$ consistent with $\mathbb{P}_{Y,T}$. As in the previous point, this is necessary for *any* estimator:

   Proposition 7. *(Lower Bound with respect to* $\mathbb{P}_{Y,T,Z}$*) Fix any confounded distribution and assume that infinite confounded data is given (or equivalently,* $\mathbb{P}_{Y,T}$ *is known). There exists* $\epsilon, \delta > 0$ *such that no* (ϵ, δ)*-close estimator exists. Specifically, for any number of deconfounded samples m,*[7] *there exist two underlying distributions* $\mathbb{P}^1_{Y,T,Z}$ *and* $\mathbb{P}^2_{Y,T,Z}$ *with the following properties:*

   - *Both of their confounded distributions are* $\mathbb{P}_{Y,T}$*.*
   - *No algorithm can correctly identify both of them with probability more than* $1 - \delta$ *using at most m deconfounded samples.*
   - *Their corresponding* ATE*'s are* ϵ *apart:* $\left|\text{ATE}(\mathbb{P}^1_{Y,T,Z}) - \text{ATE}(\mathbb{P}^2_{Y,T,Z})\right| \geq \epsilon$*.*

   The proof of Proposition 7 (§ 3.9.2) is constructive, and relies on ϵ being chosen as a function of $\mathbb{P}_{Y,T}$, and the distributions $\mathbb{P}^1_{Y,T,Z}$ and $\mathbb{P}^2_{Y,T,Z}$ being chosen as a function of δ and m. The construction relies on values of the conditional distribution $\mathbb{P}_{Z|Y,T}$ approaching 0 and 1, and so later on, we will state certain guarantees with respect to a parameter β, such that all values of $\mathbb{P}_{Z|Y,T}$ are bounded within an interval $[\beta, 1-\beta]$.

3.4 The Additional Value of Selective Deconfounding

Having established the potential value of confounded data, we turn now to the second question: *what is the additional value of selective deconfounding?* The path we will take to answering

[6] This applies when the deconfounded samples are generated according to $\mathbb{P}_{Y,T,Z}$ as in this section, or are selectively deconfounded as in the following section.

[7] See footnote 6

this question is to analyze different policies for selective deconfounding, and compare their performances via instance-dependent sample complexity lower and upper bounds, and the instance-independent worst-case sample complexity upper bounds when incorporating an infinite amount of confounded data. Incidentally, this will yield a policy which we propose for use in practice. Lastly, we extend our analysis to the scenario where we have access only to a finite amount of confounded data.

Selective Deconfounding: One important consequence of our procedure for estimating the ATE in § 3.2 is that the four conditional distributions are estimated separately: the deconfounded data is partitioned into four groups, one for each $(y, t) \in \{0, 1\}^2$, and then the quantities q_{yt}^z are estimated separately. This means that the procedure does *not* rely on the fact that the deconfounded data is drawn from the exact distribution $\mathbb{P}_{Y,T,Z}$. In particular, the draws might as well have been made directly from the conditional distributions $\mathbb{P}_{Z|Y,T}$.

Suppose now that we can draw directly from these conditional distributions $\mathbb{P}_{Z|Y,T}$. This situation may arise when the confounder is fixed (like a genetic trait) and can be observed retrospectively. This leaves us with four sample selection options, namely selecting confounded samples from the four groups, $(y, t) \in \{0, 1\}^2$, to deconfound. The problem of *selective deconfounding* formally can be stated as: given a budget for selectively deconfounded samples, how should we allocate our samples among the four groups: $(y, t) \in \{0, 1\}^2$?

Selection Policies: A *sample selection policy* knows the confounded distribution (assuming infinite confounded data), and selects the number of samples to deconfound from each of the four groups $(y, t) \in \{0, 1\}^2$. Equivalently, it is a mapping from confounded distributions $\mathbf{a}$ (recall that $a_{yt} = \mathbb{P}_{Y,T}(y, t)$) to vectors $\mathbf{x} := (x_{00}, x_{01}, x_{10}, x_{11})$, where x_{yt} indicates the *proportion* of samples allocated to each group. We will consider the following three sample selection policies:

1. **Natural (NSP):** $x_{yt} = a_{yt} = \mathbb{P}_{Y,T}(y, t)$. This is "natural" in the sense that it is equivalent to drawing directly from $\mathbb{P}_{Y,T,Z}$, as we did in the previous section. Since it can be implemented by deconfounding samples at random, it corresponds to "non-selective" deconfounding.

2. Uniform (USP): $x_{yt} = 1/4$. This splits the samples evenly across all four groups. While this policy does not depend on $\mathbf{a}$, it requires selective deconfounding to implement.[8]

3. **Outcome-weighted (OWSP):** $x_{yt} = a_{yt}/\left(2\sum_y a_{yt}\right) = \mathbb{P}_{Y|T}(y|t)/2$. This splits the samples evenly across treatment groups ($T = 0$ vs. 1), but within each treatment group, the number of samples is proportional to the outcome ($Y = 0$ vs. 1).

[8] We also observe that USP with 4X number of samples (we simply select data from all 4 categories to deconfound at each step) lower bounds the estimation error of the optimal selection policy in our problem.

While the particular form of OWSP appears to be the least intuitive, we will soon show that it provides an instance-independent guarantee when considering the worst-case conditional distributions $\mathbf{q}$.

As a sanity check, note that Theorem 6 corresponds to the sample complexity upper bound for NSP.

3.4.1 Non-existence of an Optimal Policy

Before we analyze the sample complexities of the remaining sample selection policies, we first establish that there does *not* exist a sample selection policy that is optimal. Now the notion of "optimality" here needs to be defined carefully – for example, it should be independent of the particular choice of ATE estimator, and it should apply across a *set* of underlying distributions. To that end, we introduce the following definition:

Definition 3. *Fix any $\epsilon, \delta > 0$. For any sample selection policy $\boldsymbol{x}$, any confounded distribution $\boldsymbol{a}$, and any set of conditional distributions Q, define $\mu_{\mathbf{x}}(\boldsymbol{a}, Q)$ to be the minimum number of deconfounded samples such that there exists an estimator $\widehat{\text{ATE}}$ which achieves $\mathbb{P}(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon) < \delta$ for all $\boldsymbol{q} \in Q$:*

$$\mu_{\boldsymbol{x}}(\boldsymbol{a}, Q) = \min \left\{ m \in \mathbb{N} \ : \ \exists \ \widehat{\text{ATE}} \ \textit{s.t.} \ \mathbb{P}(|\widehat{\text{ATE}} - \text{ATE}(\boldsymbol{a},\boldsymbol{q})| \geq \epsilon) < \delta\} \ \forall \ q \in Q \right\}.$$

As another sanity check, note that Theorem 6 implies an upper bound on μ_{nsp}: for all $\mathbf{a}$ and Q, $\mu_{\text{nsp}}(a, Q) \leq \sup_{q \in Q} m_{\text{nsp}}(a, q)$.

Given this definition, it is natural to call a policy $\mathbf{x}$ "optimal" for a given confounded distribution $\mathbf{a}$ and a given set of confounded distributions Q if (and only if) it achieves the minimum value $\mu_{\mathbf{x}}(\mathbf{a}, Q)$ across all policies. We can now formally state the non-existence of such an optimal policy *for any* $\mathbf{a}$:

Proposition 8. *For every confounded distribution $\boldsymbol{a}$, there exists two sets of conditional distributions Q_1 and Q_2 such that any optimal sample selection policy under $(\boldsymbol{a}, Q_1)$ is not optimal under $(\boldsymbol{a}, Q_2)$.*

The proof of Proposition 8 proceeds by construction (§ 3.9.4). Proposition 8 highlights the need to compare the proposed selection policies via various performance metrics.

3.4.2 Analysis of Policies

Returning to our three defined sampling policies, we will in this subsection show four results that each compare the policies along different metrics. Table 3.2 summarizes the comparison between the three selection policies.

	Lower Bound	Upper Bound	Worst-Case Bound	Independence
NSP	$\frac{C_1}{\beta^2} \max_{y,t} \frac{a_{yt}(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2}$	$C \max_{t,z} \frac{\sum_y a_{yt}}{\left(\sum_y a_{yt} q^z_{yt}\right)^2}$	$\frac{C}{\beta^2} \max_t \frac{1}{\sum_y a_{yt}}$	—
USP	$\frac{4C_1}{\beta^2} \max_{y,t} \frac{a^2_{yt}(\sum_y a_{y\bar{t}})^2}{(\sum_y a_{yt})^2}$	$C \max_{t,z} \frac{\sum_y 4a^2_{yt}}{(\sum_y a_{yt} q^z_{yt})^2}$	$\frac{4C}{\beta^2} \max_t \frac{\sum_y a^2_{yt}}{(\sum_y a_{yt})^2}$	—
OWSP	$\frac{2C_1}{\beta^2} \max_{y,t} \frac{a_{yt}(\sum_y a_{y\bar{t}})^2}{\sum_y a_{yt}}$	$2C \max_{t,z} \frac{(\sum_y a_{yt})^2}{(\sum_y a_{yt} q^z_{yt})^2}$	$\frac{2C}{\beta^2}$	$m_{\text{owsp}} \perp\!\!\!\perp \mathbb{P}_{Y,T}$ $M_{\text{owsp}} \perp\!\!\!\perp \mathbb{P}_{Y,T,Z}$

Table 3.2: Comparison between the instance-specific lower bound, sample complexity upper bound ($\mathbf{m}$), and the worst-case sample complexity upper bound ($\mathbf{M}$). m_{owsp} dominates m_{usp}, but neither m_{owsp} nor m_{nsp} dominates the other. Instead, Theorem 8 establishes the worst-case relative performance of μ_{owsp} and μ_{nsp}. The last column illustrates that with sufficient samples, the sample complexity upper bound guarantee for OWSP is independent of the confounded distribution $\mathbb{P}_{Y,T}$, and the worst-case sample complexity upper bound for OWSP is independent of the data generating distribution $\mathbb{P}_{Y,T,Z}$.

1. Upper Bounds on Sample Complexity: First, we provide an upper bound of the sample complexity of $\mu_{\text{usp}}(\mathbf{a}, Q)$ and $\mu_{\text{owsp}}(\mathbf{a}, Q)$ for every $\mathbf{a}$ and Q by analyzing our algorithm (analogous to Theorems 5 and 6):

Theorem 7. *Incorporating (infinite) confounded data, the estimator* ATE$(\boldsymbol{a}, \hat{\mathbf{q}})$ *is* (ϵ, δ)*-close if the number of deconfounded samples, selected under the natural selection policy (NSP) is at least:*

$$m_{\text{nsp}} := C \max_{t,z} \frac{\sum_y a_{yt}}{\left(\sum_y a_{yt} q^z_{yt}\right)^2} = C \max_{t,z} \frac{\mathbb{P}_T(t)}{\mathbb{P}_{T,Z}(t,z)^2}.$$

Under the uniform selection policy (USP):

$$m_{\text{usp}} := C \max_{t,z} \frac{4 \sum_y a^2_{yt}}{\left(\sum_y a_{yt} q^z_{yt}\right)^2} = C \max_{t,z} \frac{4 \sum_y \mathbb{P}_{Y,T}(y,t)^2}{\mathbb{P}_{T,Z}(t,z)^2}.$$

Under the outcome-weighted selection policy (OWSP):

$$m_{\text{owsp}} := C \max_{t,z} \frac{2 \left(\sum_y a_{yt}\right)^2}{\left(\sum_y a_{yt} q^z_{yt}\right)^2} = C \max_{t,z} \frac{2}{\mathbb{P}_{Z|T}(z|t)^2}.$$

Here, $C := 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$.

As a sanity check, note that the first statement in Theorem 7 is a restatement of Theorem 6, reproduced here to make comparison with the rest of the theorem easier. The proof of Theorem 7 (§ 3.9.5), which differs from the proof of Theorem 5, requires a modification to Hoeffding's inequality, which we derive to bound the sample complexity of the weighted sum of two independent random variables. Theorem 7 points to two advantages of OWSP:

1. OWSP has the nice property that m_{owsp} does not depend on $\mathbb{P}_{Y,T}$. This means that when we have sufficient data, the sample complexity upper bound guarantee of OWSP is consistent across all confounded distributions $\mathbb{P}_{Y,T}$.

2. USP is entirely dominated by OWSP: $m_{\text{usp}} \geq m_{\text{owsp}}$, since $4a_{0t}^2 + 4a_{1t}^2 - 2(a_{0t} + a_{1t})^2 = 2(a_{0t} - a_{1t})^2 \geq 0$.

We might hope for a similar result by comparing m_{owsp} with m_{nsp}, but neither strictly dominates the other, and in fact Proposition 8 rules out the possibility of finding a policy which strictly dominates all others.

2. Worst-Case Upper Bounds: Taking a slightly different tack, we might consider computing what Theorem 7 guarantees for each policy, across *all* possible values of the conditional distribution $\mathbf{q}$ – this is a reasonable re-interpretation of Theorem 7 since we do not know the value of $\mathbf{q}$ in advance. One problem with this approach is that for any confounded distribution $\mathbf{a}$, each of the three values m_{nsp}, m_{usp}, and m_{owsp}, can be made arbitrarily large (which is consistent with the lower bound in Proposition 7, which recall applies to *any* choice of estimator) by taking certain values of $\mathbf{q}$ to be close to 0 or 1. So instead of considering all possible values of $\mathbf{q}$, we will parameterize this entire analysis by a constant $\beta \in (0, 1/2)$: (Worst-Case Upper Bound) Fix $\beta \in (0, 1/2)$, and let $\mathcal{Q}_\beta := \{\mathbf{q} : q_{yt}^z \in [\beta, 1-\beta] \ \forall\ y, t, z\}$. Then,

$$M_{\text{nsp}} := \max_{\mathbf{q}\in\mathcal{Q}_\beta} m_{\text{nsp}} = \frac{C}{\beta^2} \max_t \frac{1}{\sum_y a_{yt}},$$
$$M_{\text{usp}} := \max_{\mathbf{q}\in\mathcal{Q}_\beta} m_{\text{usp}} = \frac{4C}{\beta^2} \max_t \frac{\sum_y a_{yt}^2}{(\sum_y a_{yt})^2},$$
$$M_{\text{owsp}} := \max_{\mathbf{q}\in\mathcal{Q}_\beta} m_{\text{owsp}} = \frac{2C}{\beta^2}.$$

A few observations to make on Corollary 3.4:

1. The maximum of m_{nsp}, m_{usp}, and m_{owsp} are always obtained at $q_{yt}^z = \beta$ for some t, z. This further justifies for our choice of parameterization by β.

2. M_{owsp} is independent of $\mathbb{P}_{Y,T,Z}$. This means that when data is sufficient, the worst-case sample complexity upper bound for OWSP is consistent across all data generating distributions.

3. OWSP has the lowest worst-case bound:

$$M_{\text{owsp}} \leq M_{\text{nsp}} \quad \text{and} \quad M_{\text{owsp}} \leq M_{\text{usp}}.$$

To see the first inequality, note that $\min \max_t 1/(\sum_y a_{yt})$ is achieved when $\sum_y a_{yt} = 1/2$, so $\max_t 1/(\sum_y a_{yt}) \geq 2$. To see the second inequality, note that for binary-valued Y, we have $2\sum_y a_{yt}^2 \geq \sum_y a_{yt}^2 + 2\Pi_y a_{yt} = (\sum_y a_{yt})^2$.

3. Estimator-Independent Upper Bounds: Zooming back out to our goal of comparing the three different selection policies, Theorem 7 effectively eliminated USP since it is dominated by OWSP ($m_{\text{usp}} \geq m_{\text{owsp}}$) for *every* data-generating distribution. Regarding the comparison between NSP and OWSP, Corollary 3.4 showed that OWSP dominates NSP for every confounded distribution $\mathbf{a}$, under each policy's worst-case instance of $\mathbf{q}$. To expand on the comparison between OWSP and NSP, next we provide a stronger guarantee that holds for *arbitrary* estimators. Recall from Definition 3 that $\mu_{\mathbf{x}}(\mathbf{a}, \mathcal{Q})$ is the minimum number of deconfounded samples, collected under policy $\mathbf{x}$, such that *some* estimator is (ϵ,δ)-close for all $\mathbf{q} \in \mathcal{Q}$.

The following result (Proof in § 3.9.6) establishes that $\mu_{\text{nsp}}(\mathbf{a}, \mathcal{Q})$ may be arbitrarily larger than $\mu_{\text{owsp}}(\mathbf{a}, \mathcal{Q})$, but $\mu_{\text{owsp}}(\mathbf{a}, \mathcal{Q})$ is never more than twice as large:

Theorem 8. *Fix any $\beta \in (0, 1/2)$. For any $\epsilon \in (0, 0.5 - 2\beta(1-\beta)]$, there exist confounded distributions $\boldsymbol{a}$, and $\mathcal{Q} \subset \mathcal{Q}_\beta$, such that $\mu_{\text{owsp}}(\boldsymbol{a}, \mathcal{Q})/\mu_{\text{nsp}}(\boldsymbol{a}, \mathcal{Q})$ is arbitrarily close to zero. In addition, for all $\boldsymbol{a}$ and $\mathcal{Q}$, $\mu_{\text{owsp}}(\boldsymbol{a}, \mathcal{Q}) \leq 2\mu_{\text{nsp}}(\boldsymbol{a}, \mathcal{Q})$.*

4. Lower Bounds on Sample Complexity: Finally, we show lower bounds on $\mu_{\text{nsp}}(a, \mathcal{Q}_\beta)$, $\mu_{\text{usp}}(a, \mathcal{Q}_\beta)$, and $\mu_{\text{owsp}}(a, \mathcal{Q}_\beta)$ that are analogous to Theorem 6:

Theorem 9. *(Lower Bound) Fix any $\beta \in (0, 1/2)$ and any $\boldsymbol{a}$. Then,*

$$\mu_{\text{nsp}}(a, \mathcal{Q}_\beta) \geq \frac{C_1}{\beta^2} \max_{y,t} \frac{a_{yt}(\sum_{y'} a_{y'\bar{t}})^2}{(\sum_{y'} a_{y't})^2},$$

$$\mu_{\text{usp}}(a, \mathcal{Q}_\beta) \geq \frac{C_1}{\beta^2} \max_{y,t} \frac{4a_{yt}^2(\sum_{y'} a_{y'\bar{t}})^2}{(\sum_{y'} a_{y't})^2},$$

$$\mu_{\text{owsp}}(a, \mathcal{Q}_\beta) \geq \frac{C_1}{\beta^2} \max_{y,t} \frac{2a_{yt}(\sum_{y'} a_{y'\bar{t}})^2}{\sum_{y'} a_{y't}},$$

where $\bar{t} = 1 - t$ and $C_1 \propto (k\beta - 1)^2 \ln(\delta^{-1})\epsilon^{-2}$.

The proof (§ 3.9.7) proceeds by construction. When comparing the constants C and C_1, we observe that the upper and lower bounds match in k, ϵ, and δ, demonstrating the relative tightness of our analysis.

3.4.3 Incorporating Finite Confounded Data

We have now shown that *given an infinite amount of confounded data*, OWSP outperforms the NSP in the worst case. However, in practice, the confounded data will be finite. Recall that in this case, *deconfounding* reveals the value of Z for one (initially confounded) sample, and thus we gain no additional information about $\mathbb{P}_{Y,T}$. Thus, these n confounded data provide us with an *estimate* of the confounded distribution, $\hat{\mathbb{P}}_{Y,T}(y,t)$, $\hat{a}_{yt}$. To check the robustness of OWSP, we extend our analysis to handle finite confounded data. With x_{yt} defined as above, we can derive a theorem analogous to Theorems 5-7:

Theorem 10. *Given n confounded and m deconfounded samples, with $n \geq m$, $\mathrm{ATE}_{\hat{\mathbf{a}}}(\hat{\mathbf{q}})$ is (ϵ, δ)-close when*

$$\min_{y,t,z} \frac{\left(\sum_y a_{yt} q^z_{yt}\right)^2}{\frac{1}{x_{yt}m} + \frac{(q^z_{yt})^2}{n}} = \min_{y,t,z} \left(\frac{\mathbb{P}_{T,Z}(t,z)^2}{\frac{1}{x_{yt}m} + \frac{(q^z_{yt})^2}{n}} \right) \geq 4C. \tag{3.14}$$

Here, $C := 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$.

The proof of Theorem 10 (§ 3.9.8) requires a bound we derive for the product of two independent random variables. A few results follow from Theorem 10. First, a quick calculation shows that when m is held constant, $\mathbb{P}\left(|\mathrm{ATE}_{\mathbf{a}}(\mathbf{q}) - \mathrm{ATE}_{\hat{\mathbf{a}}}(\hat{\mathbf{q}})| \geq \epsilon\right)$ remains positive as $n \longrightarrow \infty$. This means that for a certain combinations of ϵ, δ, n, there does not necessarily exist a sufficiently large m s.t. $\mathbb{P}\left(|\mathrm{ATE}_{\mathbf{a}}(\mathbf{q}) - \mathrm{ATE}_{\hat{\mathbf{a}}}(\hat{\mathbf{q}})| \geq \epsilon\right) \leq \delta$ can be satisfied. However, when there exists such an m, then $m \geq \max_{y,t,z} x^{-1}_{yt} \left(\mathbb{P}_{T,Z}(t,z)^2/(4C) - (q^z_{yt})^2/n\right)^{-1}$. Although Theorem 10 does not recover Theorem 7 exactly when $n \longrightarrow \infty$,[9] it provides insights into the relative performance of our sampling policies. Moreover, a conclusion that is similar to Theorem 8 holds: $m_{\text{owsp}}/m_{\text{nsp}} \leq 2$, and there exist distributions $\mathbb{P}_{Y,T,Z}$ such that $m_{\text{owsp}}/m_{\text{nsp}}$ is arbitrarily small. Theorem 10 also implies that when $n \gg (q^z_{yt})^2 x_{yt} m \ \forall (y,t) \in \{0,1\}^2$, the majority of the estimation error comes from not deconfounding enough data. This is because when the number of confounded data that we have is more than $\Omega(m)$, the error on the ATE in Eq. (3.14) is dominated by fact that we have enough deconfounded data. Put another way, for a given m, it is sufficient to have $n = \Omega(m)$ confounded samples.

One new issue that arises with finite confounded data is that a sampling policy may not be feasible because there are not enough confounded samples to deconfound. This does not happen for NSP (assuming $m \leq n$), but can occur for USP and OWSP. When this happens, we approximate the target sampling policy as closely as is feasible (see § 3.5.2).

3.5 Experiments

Since the upper bounds that we derived in § 3.3 are not necessarily tight, we first perform synthetic experiments to assess the tightness of our bounds. For the purpose of illustration, we focus on

[9] We could apply Lemma 5 to obtain a bound that recovers Theorem 7 exactly as $n \longrightarrow \infty$. However, this method does not give us sufficient insights into the comparative performance of our sampling policies.

binary confounders Z throughout this section, with $q_{yt} = \mathbb{P}_{Z=1|Y,T}(y,t)$ and $\mathbf{q} := (q_{00}, q_{01}, q_{10}, q_{11})$. We first compare the sampling policies in synthetic experiments on randomly chosen distributions $\mathbb{P}_{Y,T,Z}$, measuring both the average and worst-case performance of each sampling policy. In terms of average performance, we find that OWSP outperforms NSP and USP (Fig. 3.2), and there exists data generating distributions in which OWSP underperforms NSP and USP (Fig. 3.3). When averaged over the conditional distribution, OWSP outperforms both NSP and OWSP (Fig. 3.4). In addition, we numerically investigate the data generating distributions in which OWSP outperforms NSP and USP (Fig. 3.4). We discover that 1) the advantage of OWSP over NSP is the largest when the treatment group is highly inbalanced, and 2) the advantage of OWSP over USP is the largest when the outcome is highly inbalanced within each treatment group. We then measure the effect of having finite (vs. infinite) confounded data (Fig. 3.5), demonstrating that OWSP is robust under finite confounded data. Finally, we test the performance of OWSP on real-world data taken from a genetic database, COSMIC, that includes genetic mutations of cancer patients (Tate et al. 2019, Cosmic 2019) (Fig. 3.6), showing the benefit of OWSP in real-world applications. Because this is (to our knowledge) the first paper to investigate the problem of *selective deconfounding*, the methods described in § 3.1.2 are not directly comparable to ours.

3.5.1 Infinite Confounded Data: Synthetic Experiments

Assuming access to infinite confounded data, we experimentally evaluate all four sampling methods for estimating the ATE: using deconfounded data alone, and using confounded data that has been selected according to NSP, USP, and OWSP. We evaluate the performance of four methods in terms of the *absolute error*, $|\widehat{\text{ATE}} - \text{ATE}|$. Because the variance of our estimators cannot be analyzed in closed form, we report the variance of the *absolute error* averaged over different instances.

Performance on Randomly Generated Instances: We first evaluate the four methods over a randomly-selected set of distributions. Fig. 3.2 was generated by averaging over 13,000 instances, each with the distribution $\mathbb{P}_{Y,T,Z}$ drawn uniformly from the unit 7-Simplex. Every instance consists of 100 replications, each with a random draw of 1,200 deconfounded samples. The absolute error is measured as a function of the number of deconfounded samples in steps of 100 samples. Fig. 3.2 (top left) compares the use of deconfounded data along with the incorporation of confounded data selected naturally (as in the comparison of Theorems 5 and 7). It shows that incorporating confounded data yields a significant improvement in estimation error. For example, achieving an absolute error of 0.02 using deconfounded data alone requires more than 1,200 samples on average, while by incorporating confounded data, only 300 samples are required. We observe that by incorporating infinite amount of confounded dataset, the variance of our estimator has decreased dramatically. Having established the value of confounded data, Fig. 3.2 (top middle) compares the three selection policies. We find that, when averaged over joint distributions, OWSP outperforms both NSP and USP in terms of both the average absolute error and the variance. Fig. 3.2 (top right) compares the average squared error of the three selection policies. We find that OWSP outperforms NSP and USP in terms of estimation bias as well. To compare the performance of our sampling policies on an instance level, we provide three scatter plots in Fig. 3.2 (middle), each containing the 13,000 instances in the top figures and averaged over 100 replications. The number of deconfounded samples is fixed at 1,200. We observe that OWSP outperforms NSP and USP in the majority of instances. Furthermore, to compare the variance of different sampling policies on an instance level, we provide three additional scatter plots in Fig. 3.2 (bottom). At each level of deconfounded samples, all figures in the bottom row of Fig. 3.2 contain the 13,000 instance in the top figures. Each dot is calculated by taking difference between the variances of selected sampling policies, where the variance is calculated using the same 100 replications contained in the top figures. Fig. 3.2 (bottom left) contains the difference between the variance of NSP and the variance of USP on the same instances. A dot with a positive y-axis value represents that the variance of USP on a particular instance is lower than that of NSP at the given level of deconfounded data. We observe that the variance of USP is lower than NSP on the majority of instances, and vice versa if the y-axis value is negative. Similarly, a dot with a positive y-axis value on Fig. 3.2 (bottom middle and right) represents that the variance of OWSP is lower than these of NSP and USP, respectively, on a particular instance at a given level of deconfounded data. We observe that the variance of OWSP is lower than these of NSP and USP on the majority of instances.

Worst-Case Instances: In Fig. 3.3, we evaluate the performance of the three selection policies on joint distributions chosen adversarially against each. The three sub-figures (the columns) correspond to instances where NSP, USP, and OWSP perform the worst, respectively, from the left to the right. Each sub-figure is further subdivided: the top contains results for the single adversarial example while the bottom is averaged over 500 $\mathbf{q}$'s sampled uniformly from $[0, 1]^4$. The absolute error is averaged over 10,000 replications in the top figures and over 500 in the bottom. In all cases, we draw 500 deconfounded samples and measure the absolute error in steps of 50 samples. Fig. 3.3 (left) validates Corollary 8. We observe that when the distribution of $\mathbf{a}$ is heavily skewed towards $(Y = 0, T = 0)$, OWSP and USP significantly outperform NSP. Fig. 3.3 (middle) shows that USP can underperform NSP, but when averaged over all possible values of $\mathbf{q}$, USP performs better than NSP. Fig. 3.3 (right) illustrates that OWSP can underperform NSP and USP, but, when compared with the left and middle column, the performance of OWSP is close to that of NSP and USP. In Fig. 3.3 (bottom), when averaged over all possible values of $\mathbf{q}$, OWSP outperforms both. Finally, OWSP's variance is the lowest across all scenarios. § 3.10 provides representative examples in which each of these joint distributions could appear.

Insights: To better understand the properties of the confounded distributions in which OWSP performs better than its counterparts, we conduct additional experiments. Fig. 3.4 is generated with a different set of 13,000 data generating distributions than Fig. 3.2. In particular, the 13,000 distributions of $\mathbb{P}_{Y,T,Z}$ is generated with 130 different confounded distributions $\mathbf{a}$, and under each confounded distribution, we generate 100 different conditional distributions $\mathbf{q}$. Similar to Fig. 3.2, every instance consists of 100 replications, and the absolute error is measured as a function of the number of deconfounded in steps of 100 samples. Fig. 3.4 (top row) investigate the relationship between the performance of a pair of selected methods and the level of confoundedness in an instance. Specifically, Fig. 3.4 (top row) contain 13,000 dots, each representing one instance and the number of deconfounded samples is fixed at 1,200. A dot with a positive y-axis value in top left figure represents that in this particular instance, USP yields a smaller average absolute estimation error than NSP, and vice versa if the y-axis value is negative. Similarly, in the top middle and right figures, a dot with a positive y-axis value represents that OWSP yields a smaller average absolute estimation error than NSP and USP, respectively. The level of confoundedness (the y-axis) is the absolute difference between the true ATE of an instance and the "confounded ATE" ($\mathbb{P}_{Y|T}(1|1) - \mathbb{P}_{Y|T}(1|0)$), i.e., the level of confoundedness equals to $\left|\text{ATE}_{\mathbf{a}}(\mathbf{q}) - \mathbb{P}_{Y|T}(1|1) + \mathbb{P}_{Y|T}(1|0)\right|$. Fig. 3.4 (top row) demonstrate that there is no obvious association between the level of confoundedness and the relative performance of our algorithms.

Next, we investigate the association between the performance of a pair of selected methods and the level of treatment inbalance of an instance, where the latter is calculated by $|\sum_y a_{y1} - \sum_y a_{y0}|$. Fig. 3.4 (middle row left) shows that USP works better than NSP when averaged over the (initially unobserved) conditional distribution $\mathbf{q}$ on the majority of instances. In particular, the benefit of USP over NSP increases when the treatment inbalance of an instance increases. Fig. 3.4 (middle row middle and right) first validate the observation that we made in Fig. 3.3 – when averaged over the conditional distributions $\mathbf{q}$, OWSP outperforms NSP and USP on almost all instances. In addition, Fig. 3.4 (middle row middle) shows that the treatment inbalance of an instance alone explains the scenarios in which OWSP significantly outperforms NSP, i.e., when the treatment is significantly inbalanced in the observed confounded data. However, Fig. 3.4 (middle row right) shows that the benefit of OWSP over USP cannot be explained using the level of treatment inbalance alone.

Finally, we investigate the relationship between the relative performance of a pair of selected methods and the level of outcome inbalance within each treatment group in Fig. 3.4 (bottom). In particular, the level of inbalance within each treatment group is calculated by taking the maximum outcome inbalance of each treatment group, i.e., $\max_t(|a_{1t}/\sum_y a_{yt} - 0.5|)$. Fig. 3.4 (bottom right) illustrates that the benefit of OWSP over USP increases when the maximum level of outcome inbalance within each treatment group increases.

3.5.2 Finite Confounded Data

Approximate Sampling policies under finite confounded data: To deconfound according to NSP with finite confounded data is to deconfound the first m confounded data. For USP, we split the samples to the four groups as evenly as possible. That is, we max out the bottleneck group/groups and distribute the excess data as evenly as possible among the remaining groups. Under OWSP, we have $x_{yt} = \hat{a}_{yt}/\sum_y \hat{a}_{yt}$, and when implementing OWSP, we will first ensure that the deconfounded samples are split as evenly as possible across treatment groups, and then within the each group, we split the samples close as possible according to the outcome ratio.

Results: Given only n confounded data, we test the performance of the OWSP against that of NSP and USP. In Fig. 3.5, the absolute error is measured as a function of the number of confounded samples in step sizes that increment in the log scale from 100 to 10,000 while fixing the number of deconfounded samples to 100. Fig. 3.5 (left) is generated by averaging over 13,000 instances, each consisted of 100 replications, and it compares three offline sampling selection policies. Since when we only have 100 confounded samples, the three sampling policies are identical, the error curves corresponding to NSP, USP and OWSP start at the same point on the top left corner. We observe that as the number of confounded samples increases, OWSP quickly outperforms NSP and USP on average, and the gaps between OWSP and the other two selection policies widen. Since we fix the number of deconfounded samples to be 100, all three sampling policies are equivalent when there are only 100 confounded samples in the dataset (i.e., we need to deconfound all 100 confounded samples in all cases), and the average absolute errors of the three selection policies do not converge to 0 in Fig. 3.5.

Fig. 3.5 (middle) contains the 13,000 instances described above averaged over 100 replications. It compares the performance of OWSP with that of the NSP on an instance level. Similarly, Fig. 3.5 (right) compares the performance of OWSP with that of the USP. In both figures, we fix the number of confounded samples to be 681. We observe that OWSP dominates NSP and USP in the majority of instances by both the absolute error and variance. Note that if we fix the number of confounded samples and increase the number of deconfounded samples (with $m \leq n$), we observe that OWSP dominates USP and NSP when the number of deconfounded samples are small, and the gap shrinks as the number of deconfounded samples increases. When at $m = n$, all three methods are equivalent.

3.5.3 Real-World Experiments: Cancer Mutations

Data Previously, we chose the underlying distribution $\mathbb{P}_{Y,T,Z}$ uniformly from the unit 7-Simplex. However, real-world problems of interest may not be uniformly distributed. Since causal-inference methods can be hard to validate as the true causal effect is almost never observed, to illustrate the practicality of our methods, we consider a real-world observational dataset, picking three variables to be the outcome, treatment, and confounder, and artificially hiding the confounder for some examples. Finally, we evaluate our proposed sampling methods under the assumption that we have access to infinitely many confounded samples. The Catalogue Of Somatic Mutations In Cancer (COSMIC) is a public database of DNA sequences of tumor samples. It consists of targeted gene-screening panels aggregated and manually curated over 25,000 peer reviewed papers. We focus on the variables: `primary cancer site`, and `gene`. Specifically, for 1,350,015 cancer patients, we observe their type of cancer, and for a subset of genes, whether or not a mutation was observed in each gene.

Causal Models In our experiments, we designate cancer type as the outcome, a particular mutation as the treatment, and another mutation as the confounder – this setup seems reasonable because it is well known that multiple genetic mutations are correlated with individual cancer types (Knudson 2001), and that mutations can cause both cancer itself and other mutations. As a concrete example, mutations in the genes that code RNA polymerases (responsible for ensuring the accuracy of replicating RNA) are found to increase the likelihood of both other mutations and certain cancer types (Rayner et al. 2016). The setting where the treatment mutation and cancer outcome are observed and the confounding mutation is unobserved is plausible because it is common that the majority of patients only have a subset of genes sequenced (e.g., from a commercial panel). For the purpose of illustration, we assume there is no other unobserved confounders in this subsection.

The top 6 most commonly mutated genes were selected as treatment candidates. For each combination of a cancer type and one of these genes, we removed patients for whom this gene was not sequenced, and kept all pairs that had at least 40 patients in each of the four treatment-outcome groups (to ensure our deconfounding policies would have enough samples to deconfound). This procedure gave us 275 unique combinations of a cancer (outcome), mutation (treatment), and another mutation (confounder). Since on average, each {cancer, mutation, mutation} tuple contains around 25,883 patients, we took the estimated empirical distribution as the data-generating distribution and applied the ATE formula described in § 3.3 to obtain the "true" ATE. To model the unobserved confounder, we hid the values of the confounder, only revealing the value to a sampling policy when it requested a deconfounded sample. We compared the use of deconfounded data along with the incorporation of confounded data under the three sampling selection polices: NSP, USP, and OWSP.

Results: Fig. 3.6 (left) was generated with these 275 instances each repeated for 10,000 replications. The absolute error is measured as a function of the number of deconfounded samples in step sizes of 15. First, similar to Fig. 3.2, we observe that incorporating confounded data reduces both the absolute estimation error and the variance of the estimator by a large margin. Note the improvement of OWSP over NSP is larger in this case as compared to that seen in Fig. 3.2. Furthermore, when the number of deconfounded samples is small, OWSP outperforms USP. Note that Fig. 3.6 (left) does not start with 0 because absent any deconfounded data, the estimated ATE is the same for all sampling policies. In Fig. 3.6 (middle, right), we fix the number of deconfounded samples to be 45 and compare the performance of OWSP against that of NSP and USP, respectively. Both figures contain the 275 instances in the left figure, averaged over 10,000 replications. We observe that under this setup, OWSP dominates NSP in all instances, and outperforms USP in the majority of instances.

3.6 Conclusion

Although extensive studies have been conducted in causal inference, none addresses the case where revealing the value of the confounder is the only option to estimate the causal effect. In this paper, we propose the problem of causal inference with *selectively deconfounded* data, and provide a set of non-adaptive sample selection policies. Our theoretical results upper bound the amount of deconfounded data required under each sample selection policy and provide insights for why the outcome-weighted selection policy works better on average than natural selection policy. Furthermore, we conduct extensive experiments using both synthetic and real-world data to validate our theoretical results. Note that although missing data could potentially be a limiting factor to deconfounding samples in our problem setting, when the amount of confounded data is ample, we can often assume that we will be able to deconfound enough samples to derive correct causal relationships. On the other hand, if indeed our confounded data does not contain enough samples to deconfound and intervention is not feasible on the target treatment, then one is only left with collecting new deconfounded data (potentially in additional to the confounded data). Note that in this setting our method of selecting new samples to collect can still be applied. Finally, we conclude by pointing to several promising directions for potential future research:

1. In our current model, we assume that the treatment and outcome variables are binary, and the confounders are categorical. We plan to extend our results to more general causal problems, including the cases where the causal model is linear or semi-parametric.

2. Although for confounders like genetics, the only option to estimate the causal effect is to reveal the value of the confounder, in practice, proxies and mediators might be available for a subset of confounders. Thus, we may extend the idea of selective revelation of information beyond confounders to incorporate other variables, such as mediators and proxies.

3. Finally, our work can be extended to an adaptive setting where we can dynamically update the sample selection decision once more information about the conditional probability q_{yt}^{z} is revealed.

3.7 The Generalization of Our Models

3.7.1 Multiple Confounders

In this section, we show that because we do not impose any independence assumption on the set of confounder, revealing the values of all confounders offers maximal information on the joint distribution of the confounders. In particular, we will illustrate through the case where we have two binary confounders. The extension to multiple categorical confounders is straight forward.

In the case where we have two binary confounders Z_1 and Z_2, we can express the ATE as follows:

$$\text{ATE} = \sum_{z_1, z_2} \Big(P_{Y|T,Z_1,Z_2}(1|1, z_1, z_2) - P_{Y|T,Z_1,Z_2}(1|0, z_1, z_2) \Big) P_{Z_1,Z_2}(z_1, z_2).$$

With an infinite amount of confounded data, we are provided with the joint distribution $P_{Y,T}(y, t)$. Thus, it remains to estimate the conditional distributions $P_{Z_1,Z_2|Y,T}$. In our paper, we consider only the non-adaptive policies, i.e., the number of samples to deconfound in each group (y, t) is fixed a priori. In the case where the costs of revealing the values of Z_1 and Z_2 are the same and we do not have any prior knowledge on the distributions of Z_1 and Z_2, the variables Z_1 and Z_2 becomes exchangeable. In the case where the sample selection policies are completely non-adaptive (which is the case that we consider in this paper), by the symmetry of the variables Z_1 and Z_2, we have that sampling from the joint distribution of Z_1 and Z_2 yields the maximum expected information on the value of the ATE. (Note that if the confounders take categorical values of different sizes and we allow adaptive policies, then we might be able to reduce the total cost of deconfounding to estimate the ATE to within a desired accuracy level.)

3.7.2 Pretreatment Covariates

In the case where we have known pretreatment covariates X, our model can be applied in estimating the individual treatment effect where we make the common ignorability assumption on the pretreatment covariates X and the confounder Z: given pretreatment covariates X and the confounder Z, the values of outcome variable, $Y = 0$ and $Y = 1$, are independent of treatment assignment. In this case, the distributions $P_{Y,T}(y,t)$ and $P_X(x)$ are known and the Individual Treatment Effect (ITE):

$$\begin{aligned}\text{ITE} &= \sum_{z,x}\Big(P_{Y|T,Z,X}(1|1,z,x) - P_{Y|T,Z,X}(1|0,z,x)\Big)P_{Z,X}(z,x) \\ &= \sum_{z,x}\Big(P_{Y|T,Z,X}(1|1,z,x) - P_{Y|T,Z,X}(1|0,z,x)\Big)P_{Z|X}(z|x)P_X(x). \qquad (3.15)\end{aligned}$$

Note that in Eq. (3.15) the only distributions we need to estimate are the conditional distributions $P_{Z|Y,T,X}$. The values of $P_{Y|T,Z,X}$ and $P_{Z|X}$ can be calculated from $P_{Z|Y,T,X}$ by first conditioning the confounded distributions $P_{Y,T}$ on the values of the pretreatment covariates X, i.e., we first subsample all confounded (outcome, treatment) pairs for a fixed value of X, $X = x$, and then within each subsample, estimate the conditional distributions $P_{Z|Y,T,X}$ by applying our methods. To obtain ITE, we weight the estimates we obtain from all subsamples by $P_X(x)$.

3.8 Doubly Robust Estimator When Incorporating Confounded Data

When incorporating (infinite) confounded data, we know the confounded distribution, i.e., $\mathbb{P}_{Y,T}(y,t)$'s (or the a_{yt}'s) are known. Thus, to obtain the maximum likelihood estimator for the outcome regression model, we need to add the constraint indicating that the estimated $\mathbb{P}_{Y,T}(y,t)$'s, i.e., $\widehat{\mathbb{P}}_{Y,T}(y,t)$'s, should equal to the known distribution a_{yt}'t. Thus, to integrate these constraints into the estimation processes for the conditional outcomes $\mathbb{P}_{Y|T,Z}(1|t,z)$'s, we need to estimate the additional parameters $\mathbb{P}_{Z|T}(z|t)$'s. Let z_i^t to denote the values of Z when $T = t$, where $i = 1, ..., N_t$ (where N_t is the total number of samples where $T = t$, and note that $N_t = \sum_z N_{tz}$). Now, we can write the constraint as

$$\sum_z \widehat{\mathbb{P}}_{Y|T,Z}(1|t,z)\widehat{\mathbb{P}}_{Z|T}(z|t) = \frac{a_{1t}}{\sum_y a_{yt}},$$

and we will estimate the $\mathbb{P}_{Z|T}(z|t)$ through the MLE estimator using regression: $\mathbb{P}_{Z|T}(z|t) = \sigma(w_t u_t + b_t)$, where w_t and b_t are the weights and bias respectively, and u_t is a binary random variable that takes the value 1 if $T = t$ and 0 otherwise. Thus, to obtain the MLE estimator for the outcome regression model when incorporating confounded data, we need to solve the following systems of constrained MLE problem:

$$(\mathcal{P}1) \text{ for all } t, z \ :$$
$$\begin{aligned} \arg\max \sum_{i=1}^{N_{tz}} y_i^{tz} \log(\sigma(w_{tz}u_{tz} + b_{tz}) + (1 - y_i^{tz}) \log(1 - \sigma(w_{tz}u_{tz} + b_{tz})) \\ \text{s.t.} \sum_z \sigma(w_{tz}u_{tz} + b_{tz})\sigma(w_t u_t + b_t) = \frac{a_{1t}}{\sum_y a_{yt}} \end{aligned}$$
$$(\mathcal{P}2) \text{ for all } t, z \ :$$
$$\begin{aligned} \arg\max \sum_{i=1}^{N_t} z_i^t \log(\sigma(w_t u_t + b_t) + (1 - z_i^t) \log(1 - \sigma(w_t u_t + b_t)) \\ \text{s.t.} \sum_z \sigma(w_{tz}u_{tz} + b_{tz})\sigma(w_t u_t + b_t) = \frac{a_{1t}}{\sum_y a_{yt}} \end{aligned}$$

Although we can verify that our plugin estimator used in the paper (Eq. (3.6)) is feasible to the above optimization system, a direct observation from the above optimization system ($\mathcal{P}1$ and $\mathcal{P}2$) is that all decision variables that share the same treatment value, t, share the same feasible region, but their objective functions are not correlated. Thus, the above optimization system is not well-defined. We conclude that extending the doubly-robust estimator to incorporate confounded data is *not* straight-forward and requires further research.

Checking the Feasibility of Our Estimator Recall that our plugin estimator yields the following estimated $\widehat{\mathbb{P}}_{Y|T,Z}(1|t,z)$'s and $\widehat{\mathbb{P}}_{Z|T}(z|t)$'s:

$$\begin{aligned} \overline{\sigma(w_{tz}u_{tz} + b_{tz})} &= \frac{a_{1t}\hat{q}_{1t}^z}{\sum_y a_{yt}\hat{q}_{yt}^z} \\ \overline{\sigma(w_t u_t + b_t)} &= \frac{\sum_y a_{yt}\hat{q}_{yt}^z}{\sum_y a_{yt}} \end{aligned}$$

Indeed, if we plugin $\overline{\sigma(w_{tz}u_{tz} + b_{tz})}$ and $\overline{\sigma(w_t u_t + b_t)}$ in our constraints, we have

$$\sum_z \overline{\sigma(w_{tz}u_{tz} + b_{tz})\sigma(w_t u_t + b_t)} = \sum_z \frac{a_{1t}\hat{q}_{1t}^z}{\sum_y a_{yt}\hat{q}_{yt}^z} \frac{\sum_y a_{yt}\hat{q}_{yt}^z}{\sum_y a_{yt}} = \sum_z \frac{a_{1t}\hat{q}_{1t}^z}{\sum_y a_{yt}} = \frac{a_{1t}\sum_z \hat{q}_{1t}^z}{\sum_y a_{yt}} = \frac{a_{1t}}{\sum_y a_{yt}},$$

where the last equality is because $\sum_z \hat{q}_{1t}^z = \sum_z \widehat{\mathbb{P}}_{Z|Y,T}(z|1,t) = 1$.

3.9 Proofs

To begin, recall the notation introduced in § 3.3: we model the binary-valued treatment, the binary-valued outcome, and the categorical confounder as the random variables $T \in \{0,1\}$, $Y \in \{0,1\}$, and $Z \in \{1, \dots, k\}$, respectively. The underlying joint distribution of these three random variables is represented as $P_{Y,T,Z}(\cdot, \cdot, \cdot)$. To save on space for terms that are used frequently, we define the following shorthand notation:

$$p_{yt}^z = P_{Y,T,Z}(y, t, z), \quad a_{yt} = P_{Y,T}(y, t), \quad q_{yt}^z = P_{Z|Y,T}(z|y, t).$$

These terms appear frequently because, to estimate the entire joint distribution on Y, T, Z (the p_{yt}^z's), it suffices to estimate the joint distribution on Y, T (the a_{yt}'s), along with the conditional distribution of Z on Y, T (the q_{yt}^z's): $p_{yt}^z = a_{yt} q_{yt}^z$. Finally, let $\hat{p}_{yt}^z$, $\hat{a}_{yt}^z$, and $\hat{q}_{yt}^z$ be the empirical estimates of p_{yt}^z, a_{yt}^z, and q_{yt}^z, respectively, using the MLE.

3.9.1 Proof of Theorem 5

Theorem 5. *Using deconfounded data alone, the estimator* $\text{ATE}(\hat{\boldsymbol{p}})$ *as defined in Equation* (3.4) *is* (ϵ, δ)*-close if the number of deconfounded samples is at least*

$$m_{\text{base}} := C \max_{t,z} \left(\sum_y p_{yt}^z \right)^{-2} = C \max_{t,z} \frac{1}{\mathbb{P}_{T,Z}(t, z)^2},$$

where $C := 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$.

Proof. Proof of Theorem 5 This proof proceeds as follows: first, we prove a sufficient (deterministic) condition, on the errors of our estimates of p_{yt}^z's, under which $|\widehat{\text{ATE}} - \text{ATE}|$ is small. Second, we show that the errors of our estimates of p_{yt}^z's are indeed small with high probability.

Step 1: First, we can write the ATE in terms of the p_{yt}^z's as follows:

$$\text{ATE} = \sum_z \left(P_{Y|T,Z}(1|1, z) - P_{Y|T,Z}(1|0, z) \right) P_Z(z) = \sum_z \left(\left(\frac{p_{11}^z}{\sum_y p_{y1}^z} - \frac{p_{10}^z}{\sum_y p_{y0}^z} \right) \left(\sum_{y,t} p_{yt}^z \right) \right).$$

In order for the ATE to be well-defined, we assume $\sum_y p_{yt}^z \in (0, 1)$ for all t, z throughout. We can then decompose $|\widehat{\text{ATE}} - \text{ATE}|$:

$$\begin{aligned} |\widehat{\text{ATE}} - \text{ATE}| &= \left| \sum_z \left(\left(\frac{\hat{p}_{11}^z}{\sum_y \hat{p}_{y1}^z} - \frac{\hat{p}_{10}^z}{\sum_y \hat{p}_{y0}^z} \right) \left(\sum_{y,t} \hat{p}_{yt}^z \right) - \left(\frac{p_{11}^z}{\sum_y p_{y1}^z} - \frac{p_{10}^z}{\sum_y p_{y0}^z} \right) \left(\sum_{y,t} p_{yt}^z \right) \right) \right| \\ &\leq \sum_z \left| \left(\frac{\hat{p}_{11}^z}{\sum_y \hat{p}_{y1}^z} - \frac{\hat{p}_{10}^z}{\sum_y \hat{p}_{y0}^z} \right) \left(\sum_{y,t} \hat{p}_{yt}^z \right) - \left(\frac{p_{11}^z}{\sum_y p_{y1}^z} - \frac{p_{10}^z}{\sum_y p_{y0}^z} \right) \left(\sum_{y,t} p_{yt}^z \right) \right|. \end{aligned}$$

Thus, in order to upper bound $\left|\widehat{\text{ATE}} - \text{ATE}\right|$ by some ϵ, it suffices to show that

$$\left|\left(\frac{\hat{p}^z_{11}}{\sum\limits_y \hat{p}^z_{y1}} - \frac{\hat{p}^z_{10}}{\sum\limits_y \hat{p}^z_{y0}}\right)\left(\sum_{y,t} \hat{p}^z_{yt}\right) - \left(\frac{p^z_{11}}{\sum\limits_y p^z_{y1}} - \frac{p^z_{10}}{\sum\limits_y p^z_{y0}}\right)\left(\sum_{y,t} p^z_{yt}\right)\right| \le \frac{\epsilon}{k}, \quad \forall z. \tag{3.16}$$

Step 2: To bound the above terms, we first derive Lemma 5 for bounding the error of the product of two estimates in terms of their two individual errors:

Lemma 5. *For any $u, \hat{u} \in [-1, 1]$, and $v, \hat{v} \in [0, 1]$, suppose there exists $\epsilon, \theta \in (0, 1)$ such that all of the following conditions hold: (1) $|u - \hat{u}| \le (1 - \theta)\epsilon$, (2) $|v - \hat{v}| \le \theta\epsilon$, (3) $u + \epsilon \le 1$, (4) $v + \epsilon \le 1$, and (5) $\epsilon \le \min(u, v)$. Then, $|uv - \hat{u}\hat{v}| \le \epsilon$.*

Proof. Proof of Lemma 5 Since $|u - \hat{u}| \le (1 - \theta)\epsilon$, we have $\hat{u} \in [u - (1 - \theta)\epsilon, u + (1 - \theta)\epsilon]$, and similarly, from $|v - \hat{v}| \le \theta\epsilon$, we have $\hat{v} \in [v - \theta\epsilon, v + \theta\epsilon]$. Thus,

$$\begin{aligned}
|uv - \hat{u}\hat{v}| &\le \max\left(|uv - (u + (1 - \theta)\epsilon)(v + \theta\epsilon)|, |uv - (u - (1 - \theta)\epsilon)(v - \theta\epsilon)|\right) && (\text{because } v, \hat{v} \ge 0)\\
&= \max(\left|\theta u\epsilon + (1 - \theta)v\epsilon + (1 - \theta)\theta\epsilon^2\right|, \left|\theta u\epsilon + (1 - \theta)v\epsilon - (1 - \theta)\theta\epsilon^2\right|)\\
&= \left|\theta u\epsilon + (1 - \theta)v\epsilon + (1 - \theta)\theta\epsilon^2\right| && (\text{because } (1 - \theta)\theta\epsilon^2 > 0)\\
&\le |\theta(u + \epsilon)\epsilon + (1 - \theta)v\epsilon| && (\text{because } \theta\epsilon^2 > (1 - \theta)\theta\epsilon^2)\\
&\le \epsilon && (\text{because } u + \epsilon \in [-1, 1], \text{ and } v \le 1)
\end{aligned}$$

□

We can apply Lemma 5 directly to the terms in Eq. (3.16) by setting

$$u_z = \frac{p^z_{11}}{\sum\limits_y p^z_{y1}} - \frac{p^z_{10}}{\sum\limits_y p^z_{y0}}, \quad \hat{u}_z = \frac{\hat{p}^z_{11}}{\sum\limits_y \hat{p}^z_{y1}} - \frac{\hat{p}^z_{10}}{\sum\limits_y \hat{p}^z_{y0}}, \quad v_z = \sum_{y,t} p^z_{yt}, \quad \hat{v}_z = \sum_{y,t} \hat{p}^z_{yt},$$

and noting that $u_z, \hat{u}_z \in [-1, 1]$, and $v_z, \hat{v}_z \in [0, 1]$. Lemma 5 implies that the upper bound in Eq. (3.16) holds if, for some $\theta \in (0, 1)$, we have

$$|v_z - \hat{v}_z| < \frac{\theta}{k}\epsilon \quad \text{and} \quad |u_z - \hat{u}_z| < \frac{1 - \theta}{k}\epsilon.$$

While we can apply standard concentration results to the $|v_z - \hat{v}_z|$ terms, the $|u_z - \hat{u}_z|$ terms will need to be further decomposed:

$$|u_z - \hat{u}_z| = \left|\frac{p^z_{11}}{\sum\limits_y p^z_{y1}} - \frac{p^z_{10}}{\sum\limits_y p^z_{y0}} - \frac{\hat{p}^z_{11}}{\sum\limits_y \hat{p}^z_{y1}} + \frac{\hat{p}^z_{10}}{\sum\limits_y \hat{p}^z_{y0}}\right| \le \left|\frac{p^z_{11}}{\sum\limits_y p^z_{y1}} - \frac{\hat{p}^z_{11}}{\sum\limits_y \hat{p}^z_{y1}}\right| + \left|\frac{p^z_{10}}{\sum\limits_y p^z_{y0}} - \frac{\hat{p}^z_{10}}{\sum\limits_y \hat{p}^z_{y0}}\right|.$$

It will suffice to show that for each t and z,

$$\left|\frac{p^z_{1t}}{\sum\limits_y p^z_{yt}} - \frac{\hat{p}^z_{1t}}{\sum\limits_y \hat{p}^z_{yt}}\right| < \frac{1 - \theta}{2k}\epsilon. \tag{3.17}$$

Step 3: To bound these terms, we derive Lemma 6. Recall that $p^z_{1t} + p^z_{0t}, \hat{p}^z_{1t} + \hat{p}^z_{0t} \in (0, 1)$.

Lemma 6. *For any $w + s, \hat{w} + \hat{s} \in (0, 1)$, if $|w + s - \hat{w} - \hat{s}| \le (w + s)\epsilon$ and $|w - \hat{w}| \le (w + s)\epsilon$, then*

$$\left| \frac{w}{w + s} - \frac{\hat{w}}{\hat{w} + \hat{s}} \right| \le 2\epsilon.$$

Proof. Proof of Lemma 6 First, since $|w + s - \hat{w} - \hat{s}| \le (w + s)\epsilon$, we have that

$$\left| \frac{w + s}{\hat{w} + \hat{s}} - 1 \right| \le \frac{w + s}{\hat{w} + \hat{s}}\epsilon,$$

or equivalently,

$$1 - \frac{w + s}{\hat{w} + \hat{s}}\epsilon \le \frac{w + s}{\hat{w} + \hat{s}} \le 1 + \frac{w + s}{\hat{w} + \hat{s}}\epsilon.$$

We can apply this inequality and rearrange terms as follows to conclude the proof:

$$\begin{aligned}
\left| \frac{w}{w + s} - \frac{\hat{w}}{\hat{w} + \hat{s}} \right| &= \left| \frac{1}{w + s} \right| \left| w - \hat{w}\frac{w + s}{\hat{w} + \hat{s}} \right| \\
&\le \left| \frac{1}{w + s} \right| \max\left(\left| w - \hat{w}\left(1 - \frac{w + s}{\hat{w} + \hat{s}}\epsilon\right) \right|, \left| w - \hat{w}\left(1 + \frac{w + s}{\hat{w} + \hat{s}}\epsilon\right) \right| \right) \\
&= \left| \frac{1}{w + s} \right| \max\left(\left| w - \hat{w} + \frac{w + s}{\hat{w} + \hat{s}}\hat{w}\epsilon \right|, \left| w - \hat{w} - \frac{w + s}{\hat{w} + \hat{s}}\hat{w}\epsilon \right| \right) \\
&= \max\left(\left| \frac{w - \hat{w}}{w + s} + \frac{\hat{w}}{\hat{w} + \hat{s}}\epsilon \right|, \left| \frac{w - \hat{w}}{w + s} - \frac{\hat{w}}{\hat{w} + \hat{s}}\epsilon \right| \right) \\
&\le \left| \frac{w - \hat{w}}{w + s} \right| + \left| \frac{\hat{w}}{\hat{w} + \hat{s}} \right| \epsilon \le \left| \frac{w + s}{w + s} \right| \epsilon + \left| \frac{\hat{w}}{\hat{w} + \hat{s}} \right| \epsilon \le 2\epsilon.
\end{aligned}$$

The second to last inequality follows from the assumption that $|w - \hat{w}| \le (w + s)\epsilon$. □

Lemma 6 implies that Eq. (3.17) is satisfied if

$$\left| p^z_{1t} - \hat{p}^z_{1t} \right| < \frac{(\sum_y p^z_{yt})(1 - \theta)}{4k}\epsilon \quad \text{and} \quad \left| p^z_{1t} + p^z_{0t} - \hat{p}^z_{1t} - \hat{p}^z_{0t} \right| < \frac{(\sum_y p^z_{yt})(1 - \theta)}{4k}\epsilon.$$

Step 4: We've shown above that $|\widehat{\text{ATE}} - \text{ATE}| \le \epsilon$ is satisfied when

$$|v_z - \hat{v}_z| < \frac{\theta}{k}\epsilon, \quad \left| p^z_{1t} - \hat{p}^z_{1t} \right| < \frac{(\sum_y p^z_{yt})(1 - \theta)}{4k}\epsilon, \quad \text{and} \quad \left| p^z_{1t} + p^z_{0t} - \hat{p}^z_{1t} - \hat{p}^z_{0t} \right| < \frac{(\sum_y p^z_{yt})(1 - \theta)}{4k}\epsilon, \forall t, z.$$

Note that if $\forall t, \left| p^z_{1t} + p^z_{0t} - \hat{p}^z_{1t} - \hat{p}^z_{0t} \right| = \left| \sum_y p^z_{yt} - \sum_y \hat{p}^z_{yt} \right| < \frac{(\sum_y p^z_{yt})(1-\theta)}{4k}\epsilon$ then

$$|v_z - \hat{v}_z| = \left| \sum_{y,t} p^z_{yt} - \sum_{y,t} \hat{p}^z_{yt} \right| \le \sum_t \left| \sum_y p^z_{yt} - \sum_y \hat{p}^z_{yt} \right| < \frac{(\sum_{y,t} p^z_{yt})(1 - \theta)}{4k}\epsilon \le \frac{(1 - \theta)}{4k}\epsilon.$$

Thus, to remove the first constraint $|v_z - \hat{v}_z| < \frac{\theta}{k}\epsilon$, we set

$$\frac{\theta}{k}\epsilon = \frac{(1 - \theta)}{4k}\epsilon,$$

and obtain $\theta = \frac{1}{5}$.

Step 5: To summarize so far, Lemmas 5 and 6 allow us to upper bound the error of our estimated ATE in terms of upper bounds on the error of our estimates of its constituent terms:

$$P\left(|\widehat{\text{ATE}} - \text{ATE}| < \epsilon\right) \geq P\left(\bigcap_{t,z}\left\{|p^z_{1t} - \hat{p}^z_{1t}| < \frac{\sum_y p^z_{yt}}{5k}\epsilon\right\}\bigcap_{t,z}\left\{|p^z_{1t} + p^z_{0t} - \hat{p}^z_{1t} - \hat{p}^z_{0t}| < \frac{\sum_y p^z_{yt}}{5k}\epsilon\right\}\right),$$

or equivalently,

$$P\left(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon\right) \leq P\left(\bigcup_{t,z}\left\{|p^z_{1t} - \hat{p}^z_{1t}| \geq \frac{\sum_y p^z_{yt}}{5k}\epsilon\right\}\bigcup_{t,z}\left\{|p^z_{1t} + p^z_{0t} - \hat{p}^z_{1t} - \hat{p}^z_{0t}| \geq \frac{\sum_y p^z_{yt}}{5k}\epsilon\right\}\right).$$

Applying a union bound, we have

$$P\left(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon\right) \leq \sum_{t,z} P\left(|p^z_{1t} - \hat{p}^z_{1t}| \geq \frac{\sum_y p^z_{yt}}{5k}\epsilon\right) + P\left(|p^z_{1t} + p^z_{0t} - \hat{p}^z_{1t} - \hat{p}^z_{0t}| \geq \frac{\sum_y p^z_{yt}}{5k}\epsilon\right). \tag{3.18}$$

Step 6: Finally, we can apply Hoeffding's inequality (Theorem 2) to obtain the upper bound for the inequality above. Let X^z_{yt} be the random variable that maps the event $(Y = y, T = t, Z = z) \mapsto \{0, 1\}$. Then, X^z_{yt} is a Bernoulli random variable with parameter p^z_{yt}. Let m denote the total number of deconfounded samples that we have. Since $\hat{p}_{yt}$ is estimated through the MLE, we have $\hat{p}^z_{yt} = \frac{\sum_{i=1}^m X^z_{yt}}{m}$. Applying Theorem 2, we obtain:

$$P\left(\left|\frac{\sum_{i=1}^m X^z_{yt}}{m} - p^z_{yt}\right| \geq \frac{\sum_y p^z_{yt}}{5k}\epsilon\right) \leq 2\exp\left(-2m\frac{\left(\sum_y p^z_{yt}\right)^2\epsilon^2}{25k^2}\right), \quad \text{and} \tag{3.19}$$

$$P\left(\left|\frac{\sum_{i=1}^m X^z_{1t} + X^z_{0t}}{m} - p^z_{1t} - p^z_{0t}\right| \geq \frac{\sum_y p^z_{yt}}{5k}\epsilon\right) \leq 2\exp\left(-2m\frac{\left(\sum_y p^z_{yt}\right)^2\epsilon^2}{25k^2}\right). \tag{3.20}$$

Combining Eq.s (3.18), (3.19), and (3.20), we have

$$\begin{aligned}P\left(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon\right) &\leq \sum_{t,z} P\left(|p^z_{1t} - \hat{p}^z_{1t}| \geq \frac{\sum_y p^z_{yt}}{5k}\epsilon\right) + P\left(|p^z_{1t} + p^z_{0t} - \hat{p}^z_{1t} - \hat{p}^z_{0t}| \geq \frac{\sum_y p^z_{yt}}{5k}\epsilon\right)\\ &\leq 4k\max_{t,z}\left(2\exp\left(-2m\frac{\left(\sum_y p^z_{yt}\right)^2\epsilon^2}{25k^2}\right)\right) = 8k\max_{t,z}\exp\left(-2m\frac{\left(\sum_y p^z_{yt}\right)^2\epsilon^2}{25k^2}\right) \leq \delta,\end{aligned}$$

where the second line follows from the fact that, since t is binary, there are $4k$ terms in total. Solving the above equation, we conclude that $P(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon) < \delta$ is satisfied when the sample size m is at least

$$m \geq \frac{12.5k^2\ln(\frac{8k}{\delta})}{\epsilon^2}\max_{t,z}\frac{1}{\left(\sum_y p^z_{yt}\right)^2}.$$

□

3.9.2 Proof of Proposition 7

Proposition 7. *(Lower Bound with respect to $\mathbb{P}_{Y,T,Z}$) Fix any confounded distribution and assume that infinite confounded data is given (or equivalently, $\mathbb{P}_{Y,T}$ is known). There exists $\epsilon, \delta > 0$ such that no (ϵ, δ)-close estimator exists. Specifically, for any number of deconfounded samples m,*[10] *there exist two underlying distributions $\mathbb{P}^1_{Y,T,Z}$ and $\mathbb{P}^2_{Y,T,Z}$ with the following properties:*

- *Both of their confounded distributions are $\mathbb{P}_{Y,T}$.*
- *No algorithm can correctly identify both of them with probability more than $1-\delta$ using at most m deconfounded samples.*
- *Their corresponding* ATE*'s are ϵ apart: $\left|\mathrm{ATE}(\mathbb{P}^1_{Y,T,Z}) - \mathrm{ATE}(\mathbb{P}^2_{Y,T,Z})\right| \geq \epsilon$.*

Proof. Proof of Proposition 7 It suffices to show for the case where confounder takes binary value. The extension to categorical confounder is straightforward as illustrated in the proof of Theorem 9 in § 3.9.7. Let $q_{yt} = P(Z=1|Y=y, T=t)$. To show that Proposition 7 is true, it is sufficient to show that there exist a positive constant c (that depends on $\mathbf{a}$) such that for all fixed $\mathbf{a}$, there exists a pair of $\mathbf{q}$ and $\mathbf{q}'$ such that $\|\mathrm{ATE}(\mathbf{a},\mathbf{q}) - \mathrm{ATE}(\mathbf{a},\mathbf{q}')\| > c$, with $\mathbf{q}$ and $\mathbf{q}'$ close in distribution. We proceed by construction. For fixed $\mathbf{a}$, consider the following $\mathbf{q}$ pairs: $\mathbf{q} = (q_{00}, 0, q_{10}, \gamma)$ and $\mathbf{q}' = (q_{00}, \gamma, q_{10}, 0)$. Then, we have

$$\mathrm{ATE}(\mathbf{a},\mathbf{q}) = (a_{00}q_{00} + a_{10}q_{10} + a_{11}\gamma) + \frac{a_{11}(1-\gamma)}{a_{11}(1-\gamma)+a_{01}}(1 - a_{00}q_{00} - a_{10}q_{10} - a_{11}\gamma) -$$
$$\frac{a_{10}q_{10}}{a_{10}q_{10}+a_{00}q_{00}}(a_{00}q_{00} + a_{10}q_{10} + a_{11}\gamma) - \frac{a_{10}(1-q_{10})}{a_{10}(1-q_{10})+a_{00}(1-q_{00})}(1 - a_{00}q_{00} - a_{10}q_{10} - a_{11}\gamma),$$

and similarly, we have

$$\mathrm{ATE}(\mathbf{a},\mathbf{q}') = \frac{a_{11}}{a_{11}+a_{01}(1-\gamma)}(1 - a_{00}q_{00} - a_{01}\gamma - a_{10}q_{10}) - \frac{a_{10}q_{10}}{a_{10}q_{10}+a_{00}q_{00}}(a_{00}q_{00}$$
$$+ a_{01}\gamma + a_{10}q_{10}) - \frac{a_{10}(1-q_{10})}{a_{10}(1-q_{10})+a_{00}(1-q_{00})}(1 - a_{00}q_{00} - a_{01}\gamma - a_{10}q_{10}).$$

In particular,

$$\lim_{\gamma\to 0} \mathrm{ATE}(\mathbf{a},\mathbf{q}) - \mathrm{ATE}(\mathbf{a},\mathbf{q}') = a_{00}q_{00} + a_{10}q_{10} \leq a_{00} + a_{10}, \tag{3.21}$$

where we can choose q_{00} and q_{10} to be 1.

On the other hand, we can show that the number of samples needed to distinguish $\mathbf{q}$ from $\mathbf{q}'$ is at least $\Omega(1/\gamma)$: since $\mathbf{q}$ and $\mathbf{q}'$ are the same in two of the entries and symmetric on the rest two, to distinguish $\mathbf{q}$ and $\mathbf{q}'$ is to distinguish a Bernoulli random variable with parameter 0 (denoting

[10] See footnote 6

this variable B_0) from a Bernoulli random variable with parameter γ (denoting this random variable B_γ). Let f be any estimator of the Bernoulli random variable, and $x_i, ..., x_m$ be the sequence of m observations. Then we have $|\mathbb{E}_{X\sim B_0^m}[f] - \mathbb{E}_{X\sim B_\gamma^m}[f]| \le \|B_0^m - B_\gamma^m\|_1 \le \sqrt{2(\ln 2)\mathrm{KL}(B_0^m\|B_\gamma^m)} \le 2\sqrt{(\ln 2)\gamma m}$, where the last inequality is because when given m samples, $\mathrm{KL}(B_0^m\|B_\gamma^m) \le (2\gamma \ln 2 + (1-2\gamma)\ln\frac{1-2\gamma}{1-\gamma})m \le 2\gamma m$. On the other hand, any hypothesis test that takes n samples and distinguishes between $H_0 : X_1, ..., X_n \sim P_0$ and $H_1 : X_1, ..., X_n \sim P_1$ has probability of error lower bounded by $\max(P_0(1), P1(0)) \ge \frac{1}{4}e^{-n\mathrm{KL}(P_0\|P_1)}$, where $P_0(1)$ indicates the probability that we identify class H_0 while the true class is H_1. Since $P_0(1) + P_1(0) \le \delta$, by contradiction, we can show that $m \sim \Omega(\ln(\delta^{-1})\gamma^{-1})$.

Note that this lower bound on m can be arbitrarily large by choosing γ to be sufficiently small. However their ATE values stay constant away as observed in Eq. (3.21). Thus, for every fixed confounded distribution encoded by $\mathbf{a}$ and fixed number of deconfounded samples m, we can always construct a pair of conditional distributions encoded by $\mathbf{q}$ and $\mathbf{q}'$ such that their corresponding ATEs are constant away while the probability that we correctly identify the true conditional distribution from $\mathbf{q}$ and $\mathbf{q}'$ is less than $1 - \delta$. In particular, $\epsilon = c = a_{00} + a_{10}$ in the above example. (Here, we implicitly assume that $a_{00} + a_{10}$ is strictly greater than zero, i.e., $a_{00} + a_{10} > 0$.) □

3.9.3 Proof of Theorem 6

Proposition 6. *(Lower Bound with respect to ϵ and δ) Fix any confounded distribution and assume that infinite confounded data is given (or equivalently, $\mathbb{P}_{Y,T}$ is known). For any* ATE *estimator, there exists an underlying distribution $\mathbb{P}_{Y,T,Z}$, whose confounded distribution is $\mathbb{P}_{Y,T}$, for which the number of deconfounded samples*[11] *required for the estimator to be (ϵ, δ)-close is at least* $\Omega(\epsilon^{-2}\log(\delta^{-1}))$.

Proof. Proof of Theorem 6 Again, it suffices to show for the case where the confounder is binary. The extension to categorical confounder is straightforward as illustrated in the proof of Theorem 9 in § 3.9.7. Let $q_{yt} = P(Z = 1|Y = y, T = t)$. We will proceed by construction. Consider $\mathbf{q} = (q_{00}, q_{01}, \beta, \beta + \gamma)$ and $\mathbf{q}' = (q_{00}, q_{01}, \beta + \gamma, \beta)$, for some small γ. Then

$$\mathrm{ATE}(\mathbf{a}, \mathbf{q}) = \frac{a_{11}(\beta+\gamma)}{a_{11}(\beta+\gamma) + a_{01}q_{01}}(a_{00}q_{00} + a_{01}q_{01} + a_{10}\beta + a_{11}(\beta+\gamma)) + \frac{a_{11}(1-\beta-\gamma)}{a_{11}(1-\beta-\gamma) + a_{01}(1-q_{01})}$$
$$(1 - a_{00}q_{00} - a_{01}q_{01} - a_{10}\beta - a_{11}(\beta+\gamma)) - \frac{a_{10}\beta}{a_{10}\beta + a_{00}q_{00}}(a_{00}q_{00} + a_{01}q_{01} + a_{10}\beta + a_{11}(\beta+\gamma)) -$$
$$\frac{a_{10}(1-\beta)}{a_{10}(1-\beta) + a_{00}(1-q_{00})}(1 - a_{00}q_{00} - a_{01}q_{01} - a_{10}\beta - a_{11}(\beta+\gamma)),$$

[11] This applies when the deconfounded samples are generated according to $\mathbb{P}_{Y,T,Z}$ as in this section, or are selectively deconfounded as in the following section.

and similarly, we have

$$\mathrm{ATE}(\mathbf{a},\mathbf{q}') = \frac{a_{11}\beta}{a_{11}\beta + a_{01}q_{01}}(a_{00}q_{00} + a_{01}q_{01} + a_{10}(\beta+\gamma) + a_{11}\beta) + \frac{a_{11}(1-\beta)}{a_{11}(1-\beta) + a_{01}(1-q_{01})}(1 - a_{00}q_{00} - a_{01}q_{01} - a_{10}(\beta+\gamma) - a_{11}\beta) - \frac{a_{10}(\beta+\gamma)}{a_{10}(\beta+\gamma) + a_{00}q_{00}}(a_{00}q_{00} + a_{01}q_{01} + a_{10}(\beta+\gamma) + a_{11}\beta) - \frac{a_{10}(1-\beta-\gamma)}{a_{10}(1-\beta-\gamma) + a_{00}(1-q_{00})}(1 - a_{00}q_{00} - a_{01}q_{01} - a_{10}(\beta+\gamma) - a_{11}\beta).$$

Ignoring the γ in the denominator, we have that

$$\begin{aligned}
\mathrm{ATE}(\mathbf{a},\mathbf{q}) - \mathrm{ATE}(\mathbf{a},\mathbf{q}') = {} & (\frac{a_{11}}{a_{11}\beta + a_{01}q_{01}} + \frac{a_{10}}{a_{10}\beta + a_{00}q_{00}})(a_{00}q_{00} + a_{01}q_{01} + a_{10}\beta + a_{11}\beta)\gamma \\
& - (\frac{a_{11}}{a_{11}(1-\beta) + a_{01}(1-q_{01})} + \frac{a_{10}}{a_{10}(1-\beta) + a_{00}(1-q_{00})})(1 - a_{00}q_{00} - a_{01}q_{01} - a_{10}\beta - a_{11}\beta)\gamma \\
& + \frac{a_{11}^2 - a_{11}a_{10}}{a_{11}\beta + a_{01}q_{01}}\beta\gamma - \frac{a_{11}^2 - a_{11}a_{10}}{a_{11}(1-\beta) + a_{01}(1-q_{01})}(1-\beta)\gamma \\
& + \frac{a_{10}^2 - a_{11}a_{10}}{a_{10}\beta + a_{00}q_{00}}\beta\gamma - \frac{a_{10}^2 - a_{11}a_{10}}{a_{10}(1-\beta) + a_{00}(1-q_{00})}(1-\beta)\gamma \\
& + \frac{a_{11}^2}{a_{11}\beta + a_{01}q_{01}}\gamma^2 + \frac{a_{11}^2}{a_{11}(1-\beta) + a_{01}(1-q_{01})}\gamma^2 + \frac{a_{10}^2}{a_{10}\beta + a_{00}q_{00}}\gamma^2 + \frac{a_{10}^2}{a_{10}(1-\beta) + a_{00}(1-q_{00})}\gamma^2
\end{aligned} \tag{3.22}$$

Similar to the proof above, let B_1 denote the Bernoulli random variable with parameter β, and let B_2 denote the Bernoulli random variable with parameter $\beta+\gamma$. Then, given m deconfounded samples, we have $\mathrm{KL}(B_1^m \| B_2^m) \leq m\beta\ln(\frac{\beta}{\beta+\gamma}) + m(1-\beta)\ln(\frac{1-\beta}{1-\beta-\gamma}) \leq m\ln(1+\frac{\gamma}{1-\beta-\gamma}) \leq m(\frac{\gamma}{1-\beta-\gamma} - \frac{\gamma^2}{2(1-\beta-\gamma)^2})$. Thus, we have $m \sim \Omega(\frac{\ln(\delta^{-1})}{\gamma^2})$. From Eq. (3.22), we observe that $\epsilon = \|\mathrm{ATE}(\mathbf{a},\mathbf{q}) - \mathrm{ATE}(\mathbf{a},\mathbf{q}')\| \sim \Omega(\gamma)$. Combining above, we have $m \sim \Omega(\frac{\ln(\delta^{-1})}{\epsilon^2})$.

□

3.9.4 Proof of Proposition 8

Proposition 8. *For every confounded distribution* $\boldsymbol{a}$*, there exists two sets of conditional distributions* $\mathcal{Q}_1$ *and* $\mathcal{Q}_2$ *such that any optimal sample selection policy under* $(\boldsymbol{a}, \mathcal{Q}_1)$ *is not optimal under* $(\boldsymbol{a}, \mathcal{Q}_2)$*.*

Proof. Proof of Proposition 8 To show Proposition 8 holds, it suffices to construct two conditional distributions sets $\mathcal{Q}_1$ and $\mathcal{Q}_2$ such that the corresponding optimal sample selection policies differ under particular choices of $\mathbf{a}$. Similar to previous proofs, it suffices to show the case where the confounder is binary. Consider $\mathcal{Q}_1 = \{(q_{00}, \eta_2, \eta_3, \eta_4), (q_{00}, \zeta_2, \zeta_3, \zeta_4)\}$ and $\mathcal{Q}_2 = \{(\eta_1, \eta_2, \eta_3, q_{11}), (\zeta_1, \zeta_2, \zeta_3, q_{11})\}$, where the values η_i and ζ_i, $i = 1, \ldots, 4$ are known, and $\eta_i \neq \zeta_i$ for some i. Moreover, q_{00} and q_{11} represent two unknown parameters to be estimated. In particular, we will chose the values of η_i and ζ_i such that $\{\text{ATE}(\mathbf{a},\mathbf{q})\}_{\mathbf{q}\in\mathcal{Q}_1}$ are nonconstant functions of q_{00} and $\{\text{ATE}(\mathbf{a},\mathbf{q})\}_{\mathbf{q}\in\mathcal{Q}_2}$ are nonconstant functions of q_{11}.

Observe that the optimal sample selection policy under $\mathcal{Q}_1$ is $\mathbf{x}_1 = (1,0,0,0)$ while the optimal selection policy under $\mathcal{Q}_2$ is $\mathbf{x}_2 = (0,0,0,1)$. Thus, we complete the proof.

3.9.5 Proof of Theorems 6 and 7

Theorem 6. *Incorporating (infinite) confounded data, the estimator* $\text{ATE}(\boldsymbol{a},\hat{\boldsymbol{q}})$ *is* (ϵ,δ)*-close if the number of deconfounded samples is at least*

$$m_{\text{nsp}} := C \max_{t,z} \frac{\sum_y a_{yt}}{\left(\sum_y a_{yt} q_{yt}^z\right)^2} = C \max_{t,z} \frac{\mathbb{P}_T(t)}{\mathbb{P}_{T,Z}(t,z)^2}, \tag{3.13}$$

where $C := 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$.

Theorem 7. *Incorporating (infinite) confounded data, the estimator* $\text{ATE}(\boldsymbol{a},\hat{\boldsymbol{q}})$ *is* (ϵ,δ)*-close if the number of deconfounded samples, selected under the natural selection policy (NSP) is at least:*

$$m_{\text{nsp}} := C \max_{t,z} \frac{\sum_y a_{yt}}{\left(\sum_y a_{yt} q_{yt}^z\right)^2} = C \max_{t,z} \frac{\mathbb{P}_T(t)}{\mathbb{P}_{T,Z}(t,z)^2}.$$

Under the uniform selection policy (USP):

$$m_{\text{usp}} := C \max_{t,z} \frac{4\sum_y a_{yt}^2}{\left(\sum_y a_{yt} q_{yt}^z\right)^2} = C \max_{t,z} \frac{4\sum_y \mathbb{P}_{Y,T}(y,t)^2}{\mathbb{P}_{T,Z}(t,z)^2}.$$

Under the outcome-weighted selection policy (OWSP):

$$m_{\text{owsp}} := C \max_{t,z} \frac{2\left(\sum_y a_{yt}\right)^2}{\left(\sum_y a_{yt} q_{yt}^z\right)^2} = C \max_{t,z} \frac{2}{\mathbb{P}_{Z|T}(z|t)^2}.$$

Here, $C := 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$.

Proof. Proof of Theorem 7 In these theorems, we derive the concentration of the $\widehat{\text{ATE}}$ assuming infinite confounded data, and parametrize p_{yt}^z by $p_{yt}^z = a_{yt} q_{yt}^z$. Since under infinite confounded data, a_{yt}'s are known, and thus we only need to estimate the q_{yt}^z's. The key difference between Theorem 7 and Theorem 5 is that now we define the random variables X_{yt}^z to map the event $(Z = z | Y = y, T = t)$ to $\{0,1\}$. Thus, X_{yt}^z is distributed according to Bernoulli(q_{yt}^z). Thus, to decompose $\left|a_{1t}q_{1t}^z + a_{0t}q_{0t}^z - a_{1t}\hat{q}_{1t}^z - a_{0t}\hat{q}_{0t}^z\right|$, we first show the following lemma:

Lemma 7. *Let* $X_1, ..., X_{x_1 m}$ *and* $Y_1, ..., Y_{x_2 m}$ *be independent random variables in [0,1]. Then for any* $t > 0$, *we have*

$$P\left(\left|\alpha \frac{\sum_{i=1}^{x_1 m} X_i - \mathbb{E}\left[X_i\right]}{x_1 m} + \beta \frac{\sum_{j=1}^{x_2 m} Y_j - \mathbb{E}\left[Y_j\right]}{x_2 m}\right| \geq \alpha t + \beta k\right) \leq 2\exp\left(-\frac{2m(\alpha t + \beta k)^2}{\left(\frac{\alpha^2}{x_1} + \frac{\beta^2}{x_2}\right)}\right).$$

Proof. Proof of Lemma 7 First observe that

$$P\left(\alpha\frac{\sum_{i=1}^{x_1 m} X_i - \mathbb{E}\left[X_i\right]}{x_1 m} + \beta\frac{\sum_{j=1}^{x_2 m} Y_j - \mathbb{E}\left[Y_j\right]}{x_2 m} \geq \alpha t + \beta k\right)$$
$$= P\left(\frac{\alpha}{x_1}\sum_{i=1}^{x_1 m}(X_i - \mathbb{E}\left[X_i\right]) + \frac{\beta}{x_2}\sum_{j=1}^{x_2 m}(Y_j - \mathbb{E}\left[Y_j\right]) \geq m\alpha t + m\beta k\right).$$

Now, let $Z_i = \frac{\alpha}{x_1}X_i$ if $i \in [1, x_1 m]$, and $Z_i = \frac{\beta}{x_2}Y_i$ if $i \in [x_1 m + 1, (x_1 + x_2)m]$. Then applying Theorem 2, we have

$$P\left(\left|\sum_{i=1}^{(x_1+x_2)m}(Z_i - \mathbb{E}[Z_i])\right| \geq m\alpha t + m\beta k\right) \leq 2\exp\left(-\frac{2m^2(\alpha t + \beta k)^2}{\sum_{i=1}^{(x_1+x_2)m}(M_i - m_i)^2}\right) = 2\exp\left(-\frac{2m(\alpha t + \beta k)^2}{\frac{\alpha^2}{x_1} + \frac{\beta^2}{x_2}}\right).$$

■

□

As defined in § 3.3, let x_{yt} denote the percentage data we sample from the group yt.

Recall that from the proof of Theorem 5, we have

$$P\left(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon\right) \leq \sum_{t,z} P\left(\left|p_{1t}^z - \hat{p}_{1t}^z\right| \geq \frac{\sum_y p_{yt}^z}{5k}\epsilon\right) + P\left(\left|p_{1t}^z + p_{0t}^z - \hat{p}_{1t}^z - \hat{p}_{0t}^z\right| \geq \frac{\sum_y p_{yt}^z}{5k}\epsilon\right)$$
$$= \sum_{t,z} P\left(\left|a_{1t}q_{1t}^z - a_{1t}\hat{q}_{1t}^z\right| \geq \frac{\sum_y a_{yt}q_{yt}^z}{5k}\epsilon\right) + P\left(\left|a_{1t}q_{1t}^z + a_{0t}q_{0t}^z - a_{1t}\hat{q}_{1t}^z - a_{0t}\hat{q}_{0t}^z\right| \geq \frac{\sum_y a_{yt}q_{yt}^z}{5k}\epsilon\right)$$
$$= \sum_{t,z} P\left(\left|q_{1t}^z - \hat{q}_{1t}^z\right| \geq \frac{\sum_y a_{yt}q_{yt}^z}{5ka_{1t}}\epsilon\right) + P\left(\left|a_{1t}q_{1t}^z + a_{0t}q_{0t}^z - a_{1t}\hat{q}_{1t}^z - a_{0t}\hat{q}_{0t}^z\right| \geq \frac{\sum_y a_{yt}q_{yt}^z}{5k}\epsilon\right)$$
$$\leq 4k\max_{t,z}\left(2\exp\left(-2x_{1t}m\frac{\left(\sum_y a_{yt}q_{yt}^z\right)^2\epsilon^2}{25k^2a_{1t}^2}\right), 2\exp\left(-2m\frac{\left(\sum_y a_{yt}q_{yt}^z\right)^2\epsilon^2}{25k^2\sum_y\frac{a_{yt}^2}{x_{yt}}}\right)\right) \leq \delta,$$

where the second to last line follows from applying Lemma 7 to the second half of the line above it.

Solving the equation above, we have

$$m \geq \frac{12.5k^2\ln(\frac{8k}{\delta})}{\epsilon^2}\max_{t,z}\left(\frac{a_{1t}^2/x_{1t}}{\left(\sum_y a_{yt}q_{yt}^z\right)^2}, \frac{\sum_y\left(a_{yt}^2/x_{yt}\right)}{\left(\sum_y a_{yt}q_{yt}^z\right)^2}\right) = \frac{12.5k^2\ln(\frac{8k}{\delta})}{\epsilon^2}\max_{t,z}\frac{\sum_y\left(a_{yt}^2/x_{yt}\right)}{\left(\sum_y a_{yt}q_{yt}^z\right)^2}. \tag{3.23}$$

The last equality is because $a_2^2/x_2, a_1^2/x_1 > 0$. Under NSP, $x_{yt} = a_{yt}$. Thus, we have

$$m_{\text{nsp}} := \frac{12.5k^2\ln(\frac{8k}{\delta})}{\epsilon^2}\max_{t,z}\frac{\sum_y a_{yt}}{\left(\sum_y a_{yt}q_{yt}^z\right)^2}.$$

Similarly, under USP, $x_{yt} = \frac{1}{4}$, and we have

$$m_{\text{usp}} \ := \frac{12.5k^2 \ln(\frac{8k}{\delta})}{\epsilon^2} \max_{t,z} \frac{\sum_y 4a_{yt}^2}{\left(\sum_y a_{yt} q_{yt}^z\right)^2}.$$

Lastly, under OWSP, $x_{yt} = \frac{a_{yt}}{2\sum_y a_{yt}}$, and we have

$$m_{\text{owsp}} \ := \frac{12.5k^2 \ln(\frac{8k}{\delta})}{\epsilon^2} \max_{t,z} \frac{2(\sum_y a_{yt})^2}{\left(\sum_y a_{yt} q_{yt}^z\right)^2}.$$

□

3.9.6 Proof of Theorem 8

Theorem 8. *Fix any $\beta \in (0, 1/2)$. For any $\epsilon \in (0, 0.5 - 2\beta(1-\beta)]$, there exist confounded distributions $\boldsymbol{a}$, and $Q \subset Q_\beta$, such that $\mu_{\text{owsp}}(\boldsymbol{a}, Q)/\mu_{\text{nsp}}(\boldsymbol{a}, Q)$ is arbitrarily close to zero. In addition, for all $\boldsymbol{a}$ and Q, $\mu_{\text{owsp}}(\boldsymbol{a}, Q) \leq 2\mu_{\text{nsp}}(\boldsymbol{a}, Q)$.*

Proof. Proof of Theorem 8 We proceed by construction. For simplicity, we illustrate the correctness of Theorem 8 for binary confounders. The extension to the multi-valued confounder is straightforward and will be demonstrated in the proof of Theorem 9.

Consider the following example: $a_{01} = a_{10} = a_{11} = \eta$, $a_{00} = 1 - 3\eta$, and consider the following pair of $\mathbf{q}$'s: $\mathbf{q} = (\beta, \beta, \beta, c\beta)$ and $\mathbf{q}' = (\beta, \beta, \beta, \beta)$, where $c \leq \frac{1-\beta}{\beta}$ is some constant. Let $Q = \{\mathbf{q}, \mathbf{q}'\}$. Thus, to obtain an (ϵ, δ)-close estimate over Q, it suffices to distinguish $\mathbf{q}$ from $\mathbf{q}'$ with high probability. To distinguish $\mathbf{q}$ from $\mathbf{q}'$, it suffices to distinguish $c\beta$ from β, and thus the optimal selection policy under Q is to allocate all samples to the last (y,t) group. Then, we have ATE$(\mathbf{a}, \mathbf{q}) = \frac{c\beta}{1+c} + \frac{(1-c\beta)(1-\beta)}{2-c\beta-\beta} - \frac{\eta}{1-2\eta}$, and ATE$(\mathbf{a}, \mathbf{q}') = \frac{1}{2} - \frac{\eta}{1-2\eta}$. Thus, ΔATE $:= |$ATE$(\mathbf{a}, \mathbf{q}) -$ ATE$(\mathbf{a}, \mathbf{q}')|$:

$$\Delta\text{ATE} = \left| \frac{1}{2} - \frac{c\beta}{c+1} - \frac{(1-c\beta)(1-\beta)}{2-c\beta-\beta} \right|.$$

Note that when $c = \frac{1-\beta}{\beta}$, ΔATE $= 0.5 - 2\beta(1-\beta) \approx 0.5$. Thus, for any $\epsilon \in [0, 0.5 - 2\beta(1-\beta)]$, there exists some c such that $\epsilon = \Delta$ATE. Then, for any δ, let $\mu(Q)$ denote the minimum expected number of samples that we need to distinguish $\mathbf{q}$ from $\mathbf{q}'$ under the best estimator and under the optimal selection policy (described above). Note that since we only need to distinguish $c\beta$ from β, this Then under NSP, the minimum number of samples that we need under the best estimator equals to $\mu_{\text{nsp}}(\mathbf{a}, Q) \ := \mu(Q)/\eta$, and under OWSP, the minimum number of samples that we need under the best estimator equals to $\mu_{\text{oswp}}(\mathbf{a}, Q) = 4\mu$. (Note that $\mathbf{x}_{yt} = (\frac{1-3\eta}{2(1-2\eta)}, \frac{1}{4}, \frac{\eta}{2(1-2\eta)}, \frac{1}{4})$ under OWSP in this example.) Thus, $\mu_{\text{owsp}}(\mathbf{a}, Q)/\mu_{\text{nsp}}(\mathbf{a}, Q) = 4\eta$. Since in this example, η is at most 1/4, $\mu_{\text{owsp}}(\mathbf{a}, Q)/\mu_{\text{nsp}}(\mathbf{a}, Q) \leq 1$ and can be arbitrarily close to 0 as $\eta \longrightarrow 0$. (Intuitively, the first statement is true because when $\sum_t a_{0t} \ll \sum_t a_{1t}$ and $a_{00} \approx a_{01}$, it is equally important to estimate q^z_{0t}'s and q^z_{1t}'s according to the ATE expression. However, under this setup, the number of samples allocated to groups $(0, t)$'s decreases as $a_{0,t}$'s approach to 0 under NSP, while under OWSP, half of the deconfounded samples are always dedicated to estimate the q^z_{0t}'s.)

Next, we show the last sentence in Theorem 8 is true. For any fixed $\epsilon, \delta < 1$ and for any Q, we note that when $w_{\text{owsp}} := 2\mu_{\text{nsp}}(\mathbf{a}, Q) \max_t \sum_y a_{yt}$ also achieves $P(|\widehat{\text{ATE}} - \text{ATE}| \geq \epsilon) < \delta$ under the outcome-weighted selection policy. The reason is that when using w_{owsp} number of deconfounded samples, the number of deconfounded data allocated to each yt group is at least as much as those under the natural selection policy. Thus, we have $\mu_{\text{owsp}}(\mathbf{a}, Q) \leq w_{\text{owsp}} \leq 2\mu_{\text{nsp}}(\mathbf{a}, Q)$, where the last inequality is because $\max_t \sum_y a_{yt} < 1$.

□

3.9.7 Proof of Theorem 9

Theorem 9. *(Lower Bound) Fix any $\beta \in (0, 1/2)$ and any $\boldsymbol{a}$. Then,*

$$\mu_{\text{nsp}}(a, Q_\beta) \geq \frac{C_1}{\beta^2} \max_{y,t} \frac{a_{yt}(\sum_{y'} a_{y'\bar{t}})^2}{(\sum_{y'} a_{y't})^2},$$

$$\mu_{\text{usp}}(a, Q_\beta) \geq \frac{C_1}{\beta^2} \max_{y,t} \frac{4a^2_{yt}(\sum_{y'} a_{y'\bar{t}})^2}{(\sum_{y'} a_{y't})^2},$$

$$\mu_{\text{owsp}}(a, Q_\beta) \geq \frac{C_1}{\beta^2} \max_{y,t} \frac{2a_{yt}(\sum_{y'} a_{y'\bar{t}})^2}{\sum_{y'} a_{y't}},$$

where $\bar{t} = 1 - t$ and $C_1 \propto (k\beta - 1)^2 \ln(\delta^{-1})\epsilon^{-2}$.

Proof. Consider $\mathbf{q} = (q^z_{00}, q^z_{01}, q^z_{10}, q^z_{11})$ where $q^1_{01} = \beta$, $q^1_{11} = \beta + \gamma$, and $q^z_{11} = q^z_{01} - \gamma/(k-1)$ for $z = 2, ..., k$, with $\sum_z q^z_{01} = \sum_z q^z_{11} = 1$. We assume that $q^z_{11}, q^z_{01} \in [\beta, 1 - \beta]$ for some suitable β and γ for all values of Z. Similarly, we consider the $\mathbf{q}'$ where the entries of q^z_{01} and q^z_{11} are flipped, i.e., $\mathbf{q}' = (q^z_{00}, q^z_{11}, q^z_{10}, q^z_{01})$, for some small γ, where the q^z_{yt}'s are defined above. Then,

$$\begin{aligned}
&\text{ATE}(\mathbf{a}, \mathbf{q}) = \sum_z \left(\left(\frac{a_{11} q^z_{11}}{\sum_y a_{y1} q^z_{y1}} - \frac{a_{10} q^z_{10}}{\sum_y a_{y0} q^z_{y0}} \right) \sum_{y,t} a_{yt} q^z_{yt} \right) \\
&= \frac{a_{11}(\beta + \gamma)}{a_{11}(\beta + \gamma) + a_{01}\beta}(a_{00} q^1_{00} + a_{01}\beta + a_{10} q^1_{10} + a_{11}(\beta + \gamma)) - \frac{a_{10} q^1_{10}}{a_{10} q^1_{10} + a_{00} q^1_{00}}(a_{00} q^1_{00} + a_{01}\beta + a_{10} q^1_{10} + \\
&a_{11}(\beta + \gamma)) + \sum_{z=2}^{k} \frac{a_{11}\left(q^z_{01} - \frac{\gamma}{k-1}\right)}{a_{11}\left(q^z_{01} - \frac{\gamma}{k-1}\right) + a_{01} q^z_{01}} \left(a_{00} q^z_{00} + a_{01} q^z_{01} + a_{10} q^z_{10} + a_{11}\left(q^z_{01} - \frac{\gamma}{k-1} \right) \right) - \\
&\sum_{z=2}^{k} \frac{a_{10} q^z_{10}}{a_{10} q^z_{10} + a_{00} q^z_{00}} \left(a_{00} q^z_{00} + a_{01} q^z_{01} + a_{10} q^z_{10} + a_{11}\left(q^z_{01} - \frac{\gamma}{k-1} \right) \right),
\end{aligned}$$

and similarly, we have

$$\mathrm{ATE}(\mathbf{a}, \mathbf{q}') = \frac{a_{11}\beta}{a_{11}\beta + a_{01}(\beta + \gamma)}(a_{00}q_{00}^1 + a_{01}(\beta + \gamma) + a_{10}q_{10}^1 + a_{11}\beta) - \frac{a_{10}q_{10}^1}{a_{10}q_{10}^1 + a_{00}q_{00}^1}(a_{00}q_{00}^1 + a_{01}(\beta + \gamma) + a_{10}q_{10}^1 + a_{11}\beta) + \sum_{z=2}^{k} \frac{a_{11}q_{01}^z}{a_{11}q_{01}^z + a_{01}\left(q_{01}^z - \frac{\gamma}{k-1}\right)} \left(a_{00}q_{00}^z + a_{01}\left(q_{01}^z - \frac{\gamma}{k-1}\right) + a_{10}q_{10}^z + a_{11}q_{01}^z\right) - \sum_{z=2}^{k} \frac{a_{10}q_{10}^z}{a_{10}q_{10}^z + a_{00}q_{00}^z} \left(a_{00}q_{00}^z + a_{01}\left(q_{01}^z - \frac{\gamma}{k-1}\right) + a_{10}q_{10}^z + a_{11}q_{01}^z\right)$$

Ignoring the γ in the denominator, we have that

$$\begin{aligned}
\mathrm{ATE}(\mathbf{a}, \mathbf{q}) - \mathrm{ATE}(\mathbf{a}, \mathbf{q}') &\approx \frac{a_{11}}{a_{11}\beta + a_{01}\beta}(a_{00}q_{00}^1 + a_{01}\beta + a_{10}q_{10}^1 + a_{11}\beta)\gamma + \frac{a_{10}q_{10}^1(a_{01} - a_{11})}{a_{10}q_{10}^1 + a_{00}q_{00}^1}\gamma \\
&- \left(\sum_{z=2}^{k}\left(\frac{a_{11}/k-1}{a_{11}q_{01}^z + a_{01}q_{01}^z}(a_{00}q_{00}^z + a_{10}q_{10}^z + (a_{01} + a_{11})q_{10}^z)\right)\right)\gamma - \sum_{z=2}^{k}\frac{a_{10}q_{10}^z(a_{01} - a_{11})}{a_{10}q_{10}^z + a_{00}q_{00}^z}\frac{1}{k-1}\gamma \\
&+ \frac{a_{11}^2}{a_{11}\beta + a_{01}\beta}\gamma^2 + \sum_{z=2}^{k}\frac{a_{11}^2}{a_{11}q_{01}^z + a_{01}q_{01}^z}\frac{\gamma^2}{(k-1)^2} \\
&= \frac{a_{11}}{a_{11}\beta + a_{01}\beta}(a_{00}q_{00}^1 + a_{10}q_{10}^1)\gamma + \frac{a_{10}q_{10}^1(a_{01} - a_{11})}{a_{10}q_{10}^1 + a_{00}q_{00}^1}\gamma - \frac{a_{11}}{k-1}\sum_{z=2}^{k}\left(\frac{a_{00}q_{00}^z + a_{10}q_{10}^z}{a_{11}q_{01}^z + a_{01}q_{01}^z}\right)\gamma \\
&- \frac{1}{k-1}\sum_{z=2}^{k}\frac{a_{10}q_{10}^z(a_{01} - a_{11})}{a_{10}q_{10}^z + a_{00}q_{00}^z}\gamma + \frac{a_{11}^2}{a_{11}\beta + a_{01}\beta}\gamma^2 + \sum_{z=2}^{k}\frac{a_{11}^2}{a_{11}q_{01}^z + a_{01}q_{01}^z}\frac{\gamma^2}{(k-1)^2} \qquad (3.24)
\end{aligned}$$

Since the second order terms in γ is dominated by the first order terms in γ, thus to find the highest lower bound for sample complexity in this instance is to find the largest coefficient in front of γ.

Assuming that $\beta \ll k$ and $k\beta < 1$, then the maximum of Eq. (3.24) is achieved when $q_{00}^z = q_{10}^z = \beta$, $q_{00}^1 = q_{10}^1 = 1 - k\beta$, and $q_{01}^z = (1 - \beta)/(k - 1)$, and the coefficient in front of γ is

$$\frac{a_{11}}{a_{11} + a_{01}}(a_{00} + a_{10})(\frac{1}{\beta} - \frac{k - \beta}{1 - \beta}) \approx \frac{a_{11}}{a_{11} + a_{01}}(a_{00} + a_{10})\left(\frac{1}{\beta} - k\right).$$

Similar to the proof of Theorem 6, we have $m \sim \Omega(\frac{\ln(\delta^{-1})}{\gamma^2})$. From Eq. (3.22), we observe that $\epsilon = \|\mathrm{ATE}(\mathbf{a}, \mathbf{q}) - \mathrm{ATE}(\mathbf{a}, \mathbf{q}')\| \sim \Omega(\gamma)$. Combining above, we have $m \sim \Omega(\frac{\ln(\delta^{-1})}{\epsilon^2})$. In the case above, $\epsilon \approx \frac{a_{11}}{a_{11}+a_{01}}(a_{00} + a_{10})\frac{1}{\beta}\gamma$, thus, the number of deconfounded samples needed is approximately

$$m \propto \frac{\ln(\delta^{-1})a_{11}^2(a_{00} + a_{10})^2}{\epsilon^2(a_{11} + a_{01})^2}\left(\frac{1}{\beta} - k\right)^2.$$

Let $C_1 \propto (k\beta - 1)^2\ln(\delta^{-1})\epsilon^{-2}$. Then $m \sim \Omega\left(\frac{C_1}{\beta^2}\frac{a_{11}^2(a_{00}+a_{10}^2)}{(a_{11}+a_{01})^2}\right)$.

If we flip the values of q_{01}^z and q_{11}^z with the values of q_{00}^z and q_{10}^z in both $\mathbf{q}$ and $\mathbf{q}'$, then we have $m \sim \frac{C_1}{\beta^2}\frac{a_{10}^2(a_{01}+a_{11})^2}{(a_{10}+a_{00})^2}$. Note that this is because that the estimation error on ATE and $1 -$ ATE is symmetric. In addition, under natural selection policy, we need at least $\frac{m}{a_{11}}$ samples; uniform selection policy, we need at least $4m$ deconfounded samples; under outcome-weighted selection policy, we need at least $2\frac{a_{11}+a_{01}}{a_{11}}m$ deconfounded samples. Combining all of the above, we obtained Theorem 9.

□

3.9.8 Proof of Theorem 10

Theorem 10. *Given n confounded and m deconfounded samples, with $n \geq m$, $\mathrm{ATE}_{\hat{\boldsymbol{a}}}(\hat{\boldsymbol{q}})$ is (ϵ, δ)-close when*

$$\min_{y,t,z} \frac{\left(\sum_y a_{yt} q_{yt}^z\right)^2}{\frac{1}{x_{yt}m} + \frac{(q_{yt}^z)^2}{n}} = \min_{y,t,z} \left(\frac{\mathbb{P}_{T,Z}(t,z)^2}{\frac{1}{x_{yt}m} + \frac{(q_{yt}^z)^2}{n}} \right) \geq 4C. \tag{3.14}$$

Here, $C := 12.5k^2 \ln(8k/\delta)\epsilon^{-2}$.

Proof. Proof of Theorem 10 In this theorem, we derive the concentration for the $\widehat{\mathrm{ATE}}$ under finite confounded data. The difference between Theorem 7 and Theorem 10 is that now we need to estimate a_{yt} in addition to q_{yt}^z. Thus, to decompose $|a_{yt}q_{yt}^z - \hat{a}_{yt}\hat{q}_{yt}^z|$, we first derive Lemma 8.

Lemma 8

Lemma 8. *Let $X_1, ..., X_n$ and $Y_1, ..., Y_m$ be two sequences of Bernoulli random variables independently drawn from distribution p_1 and p_2, respectively. Let $S_X = \sum_{i=1}^{n} X_i$, $S_Y = \sum_{i=1}^{m} Y_i$. Then,*

$$P\Big(\Big|S_X S_Y - \mathbb{E}\left[S_X\right]\mathbb{E}\left[S_Y\right]\Big| \geq nmt\Big) \leq 2\exp\left(\frac{-2t^2}{\frac{1}{m} + \frac{p_2^2}{n}}\right).$$

Proof. Proof of Lemma 8 The proof follows the proof of Hoeffding's inequality:

$$
\begin{aligned}
&P\Big(S_X S_Y - \mathbb{E}[S_X]\mathbb{E}[S_Y] \geq nmt\Big) = P\Big(\exp(aS_XS_Y - a\mathbb{E}[S_X]\mathbb{E}[S_Y])) \geq \exp(anmt)\Big) && \text{(3.25)}\\
&\leq \exp(-anmt)\mathbb{E}\left[\exp(aS_XS_Y - a\mathbb{E}[S_X]\mathbb{E}[S_Y]))\right], \qquad \text{(because of Markov's inequality)} && \text{(3.26)}\\
&= \exp(-anmt)\mathbb{E}\left[\exp(aS_X(S_Y - \mathbb{E}[S_Y]) + a\mathbb{E}[S_Y](S_X - \mathbb{E}[S_X])\right]\\
&\leq \exp(-anmt)\mathbb{E}\left[\exp(a\max(S_X)(S_Y - \mathbb{E}[S_Y]) + a\mathbb{E}[S_Y](S_X - \mathbb{E}[S_X]))\right] \qquad \text{(because } S_X \geq 0\text{)} && \text{(3.27)}\\
&= \exp(-anmt)\mathbb{E}\left[\exp(an(S_Y - \mathbb{E}[S_Y]) + a\mathbb{E}[S_Y](S_X - \mathbb{E}[S_X]))\right]\\
&= \exp\left(-anmt\right)\mathbb{E}\left[\exp\left(an(S_Y - \mathbb{E}[S_Y])\right)\right]\mathbb{E}\left[\exp(a\mathbb{E}[S_Y](S_X - \mathbb{E}[S_X]))\right] \qquad \text{(because} X \perp\!\!\!\perp Y\text{)} && \text{(3.28)}\\
&= \exp(-anmt)\prod_{i=1}^{m}\prod_{j=1}^{n}\mathbb{E}\left[\exp(an(Y_i - \mathbb{E}[Y_i]))\right]\mathbb{E}\left[\exp(a\mathbb{E}[S_Y](X_j - \mathbb{E}[X_j]))\right]\\
&\leq \exp(-anmt)\prod_{i=1}^{m}\exp\left(\frac{a^2}{8}n^2\right)\prod_{j=1}^{n}\exp\left(\frac{a^2}{8}\mathbb{E}[S_Y]^2\right) && \text{(3.29)}\\
&= \exp\left(-anmt + \frac{a^2}{8}mn^2 + \frac{a^2}{8}nm^2p_2^2\right) \qquad \text{(because the minimum is achieved at } a = \frac{4t}{n + mp_2^2}\text{)} && \text{(3.30)}\\
&\leq \exp\left(-\frac{2mnt^2}{n + mp_2^2}\right) = \exp\left(-\frac{2t^2}{\frac{1}{m} + \frac{p_2^2}{n}}\right).
\end{aligned}
$$

Line (3.29) is because $Y_i - \mathbb{E}[Y_i] \in \{-\mathbb{E}[Y_i], 1-\mathbb{E}[Y_i])$, and thus $n(Y_i - \mathbb{E}(Y_i)) \in [-n\mathbb{E}[Y_i], n(1-\mathbb{E}[Y_i])]$. Furthermore, $\mathbb{E}[S_Y](X_i - \mathbb{E}[X_i]) \in (-\mathbb{E}[X]\mathbb{E}[S_Y], (1-\mathbb{E}[X])\mathbb{E}[S_Y])$. Finally, applying Hoeffding's Lemma (Lemma 1), we obtain line (3.29). □

Now we are ready to prove Theorem 10.

Proof of Theorem 10

In this theorem, we assume that the number of confounded data is finite. Thus, instead of a_{yt}, we have estimates of them, namely $\hat{a}_{yt}$. Let n_{yt} denote the number of samples in the confounded data such that $(Y = y, T = t)$. Let m^z_{yt} be the number of samples in the deconfounded data such that $(Y = y, T = t, Z = z)$. Furthermore, let $n = \sum_{y,t} n_{yt}$, $m = \sum_{y,t,z} m^z_{yt}$. Then, under our setup, we estimate a_{yt} and q^z_{yt} as follows: $\hat{a}_{yt} = \frac{n_{yt}}{n}$, and $\hat{q}^z_{yt} = \frac{m^z_{yt}}{\sum_z m^z_{yt}}$. Thus, following the proof of Theorem 5, we have

$$P\left(|\widehat{\mathrm{ATE}} - \mathrm{ATE}| < \epsilon\right) \geq P\left(\bigcap_{t,z}\left\{\left|p^z_{1t} - \hat{p}^z_{1t}\right| < \frac{\sum_y p^z_{yt}}{5k}\epsilon\right\}\bigcap_{t,z}\left\{\left|p^z_{1t} + p^z_{0t} - \hat{p}^z_{1t} - \hat{p}^z_{0t}\right| < \frac{\sum_y p^z_{yt}}{5k}\epsilon\right\}\right)$$
$$= P\left(\bigcap_{t,z}\left\{\left|a_{1t}q^z_{1t} - \hat{a}_{1t}\hat{q}^z_{1t}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{5k}\epsilon\right\}\bigcap_{t,z}\left\{\left|a_{1t}q^z_{1t} + a_{0t}q^z_{0t} - \hat{a}_{1t}\hat{q}^z_{1t} - \hat{a}_{0t}\hat{q}^z_{0t}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{5k}\epsilon\right\}\right).$$

Notice that $\left|a_{1t}q^z_{1t} + a_{0t}q^z_{0t} - \hat{a}_{1t}\hat{q}^z_{1t} - \hat{a}_{0t}\hat{q}^z_{0t}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{5k}\epsilon$ is satisfied when both

$$\left|a_{1t}q^z_{1t} - \hat{a}_{1t}\hat{q}^z_{1t}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{10k}\epsilon, \text{ and } \left|a_{0t}q^z_{0t} - \hat{a}_{0t}\hat{q}^z_{0t}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{10k}\epsilon.$$

We have:

$$P\left(|\widehat{\mathrm{ATE}} - \mathrm{ATE}| < \epsilon\right) \geq P\left(\bigcap_{t,z}\left\{\left|a_{1t}q^z_{1t} - \hat{a}_{1t}\hat{q}^z_{1t}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{10k}\epsilon\right\}\bigcap_{t,z}\left\{\left|a_{0t}q^z_{0t} - \hat{a}_{0t}\hat{q}^z_{0t}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{10k}\epsilon\right\}\right)$$
$$= P\left(\bigcap_{y,t,z}\left\{\left|a_{yt}q^z_{yt} - \hat{a}_{yt}\hat{q}^z_{yt}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{10k}\epsilon\right\}\right).$$

Lemma 8 suggests that

$$P(|a_{yt}q^z_{yt} - \hat{a}_{yt}\hat{q}^z_{yt}| \geq t) \leq 2\exp\left(-\frac{2t^2}{\frac{1}{x_{yt}m} + \frac{(q^z_{yt})^2}{n}}\right).$$

Thus, applying a union bound and Lemma 8, we have

$$P\left(|\widehat{\mathrm{ATE}} - \mathrm{ATE}| \geq \epsilon\right) \leq \sum_{y,t,z} P\left(\left|a_{yt}q^z_{yt} - \hat{a}_{yt}\hat{q}^z_{yt}\right| < \frac{\sum_y a_{yt}q^z_{yt}}{10k}\epsilon\right) \leq 8k\max_{y,t,z}\exp\left(-2\frac{\left(\sum_y a_{yt}q^z_{yt}\right)^2\epsilon^2}{(\frac{1}{x_{yt}m} + \frac{(q^z_{yt})^2}{n})100k^2}\right) \leq \delta.$$

Simplifying the equations above, we have

$$\min_{y,t,z}\frac{\left(\sum_y a_{yt}q^z_{yt}\right)^2}{(\frac{1}{x_{yt}m} + \frac{(q^z_{yt})^2}{n})} \geq \frac{50k^2\ln\left(\frac{8k}{\delta}\right)}{\epsilon^2}.$$

□

3.10 Corresponding Stories

In this section, we will provide an example for each selection method such that this particular sampling performs the worst when compared with the other two methods. For the purpose of illustration, we consider binary confounder throughout this section. To ease notation, let q_{yt} denote q^1_{yt}.

A Scenario in Which NSP Performs the Worst A drug repositioning start-up discovered that drug T can potentially cure a disease γ. which has no known drug cure and goes away without treatments once a while. Since drug T is commonly used to treat another disease η, the majority patients who has disease γ do not receive any treatment. Among the ones who received drug T, the start-up discovered that the health outcomes of the majority of patients have improved. The start-up proposes to bring drug T to an observational study to verify whether drug T could treat disease γ while not controlling for patient's treatment adherence levels. As in most cases, patient's treatment adherence levels could influence doctors' decision of whether to prescribe drug T and whether the treatment for disease γ will be successful. Translating this scenario into our notations, we have $a_{01} = \epsilon_1$, $a_{10} = \epsilon_2$, $a_{11} = \epsilon_3$, and $a_{00} = 1 - \sum_{i=1}^{3} \epsilon_i$, say $\mathbf{a} = (0.9, 0.02, 0.01, 0.07)$. Now, imagine in the clinical trial, the patients are given a drug case containing drug T such that the drug case automatically records the frequency that the patient takes the drug. Somehow we know a priori that the patients who do not have health improvement have on average poor treatment adherence, e.g., $q_{00} = 0.9$, $q_{01} = 0.7$; furthermore, those who have health improvement on average have good treatment adherence, e.g., $q_{10} = 0.01$, $q_{11} = 0.3$. Deconfounding according to NSP, i.e., $\mathbf{x} = (a_{00}, a_{01}, a_{10}, a_{11})$, in this case, will select most samples from the group $(Y = 0, T = 0)$. Since the ATE depends on the estimation that relies on both $T = 0$, and $T = 1$, one would expect that NSP and OWSP will outperform NSP. The left column in Fig. 3.3 confirms this hypothesis.

A Scenario in Which USP Performs the Worst A group biostatisticians discovered that mutations on gene T is likely to cause cancer Y in patients with a particular type of heart disease. In particular, they discovered that among the those heart disease patients, 79% of patients have neither mutation on T nor cancer Y; 18% patients have both mutation on T and cancer Y. In other words, $a_{00} = 0.79$, $a_{11} = 0.18$. Furthermore, we have $a_{01} = 0.01$, $a_{10} = 0.02$. This group of biostatisticians want to run a small experiment to confirm whether gene T causes cancer Y. In particular, they are interested in knowing whether those patients also have mutations on gene Z, which is also suspected by the same group of biostatisticians to cause cancer Y. Somehow, we know a priori that $q_{00} = 0.5$, $q_{01} = 0.01$, $q_{10} = 0.05$, $q_{11} = 0.5$. From the calculation of the ATE, it is not difficult to observe that the error on the ATE is dominated by the estimation errors on q_{00}, q_{11}. Thus, we should sample more from the groups $(Y = 0, T = 0)$ and $(Y = 1, T = 1)$.

A Scenario in Which OWSP Performs the Worst A team wants to reposition drug T to cure diabetes. Drug T has been used to treat a common comorbid condition of diabetes that appears in 31% of the diabetic patient population. Among those patients who receive drug T, about 97% has improved health, that is $a_{01} = 0.01$ and $a_{11} = 0.3$. Among the patients who have never received drug T, about 70% have no health improvement, that is $a_{00} = 0.5$, and $a_{10} = 0.19$.

Let $q_{00} = 0.05, q_{01} = 0.5, q_{10} = 0.055,$ and $q_{11} = 0.4$. In the ATE, it is easy to observe that $\frac{a_{11}q_{11}}{a_{11}q_{11}+a_{01}q_{01}}$ and $\frac{a_{11}(1-q_{11})}{a_{11}(1-q_{11})+a_{01}(1-q_{01})}$ are both dominated by 1 regardless of the estimates of q_{11} and q_{01}. In this case, USP outperforms OWSP and NSP when the sample size is larger than 200. On the other hand, the bottom figure in the third column of Fig. 3.3 shows that, when averaged over all possible values of $\mathbf{q}$, OWSP performs the best.

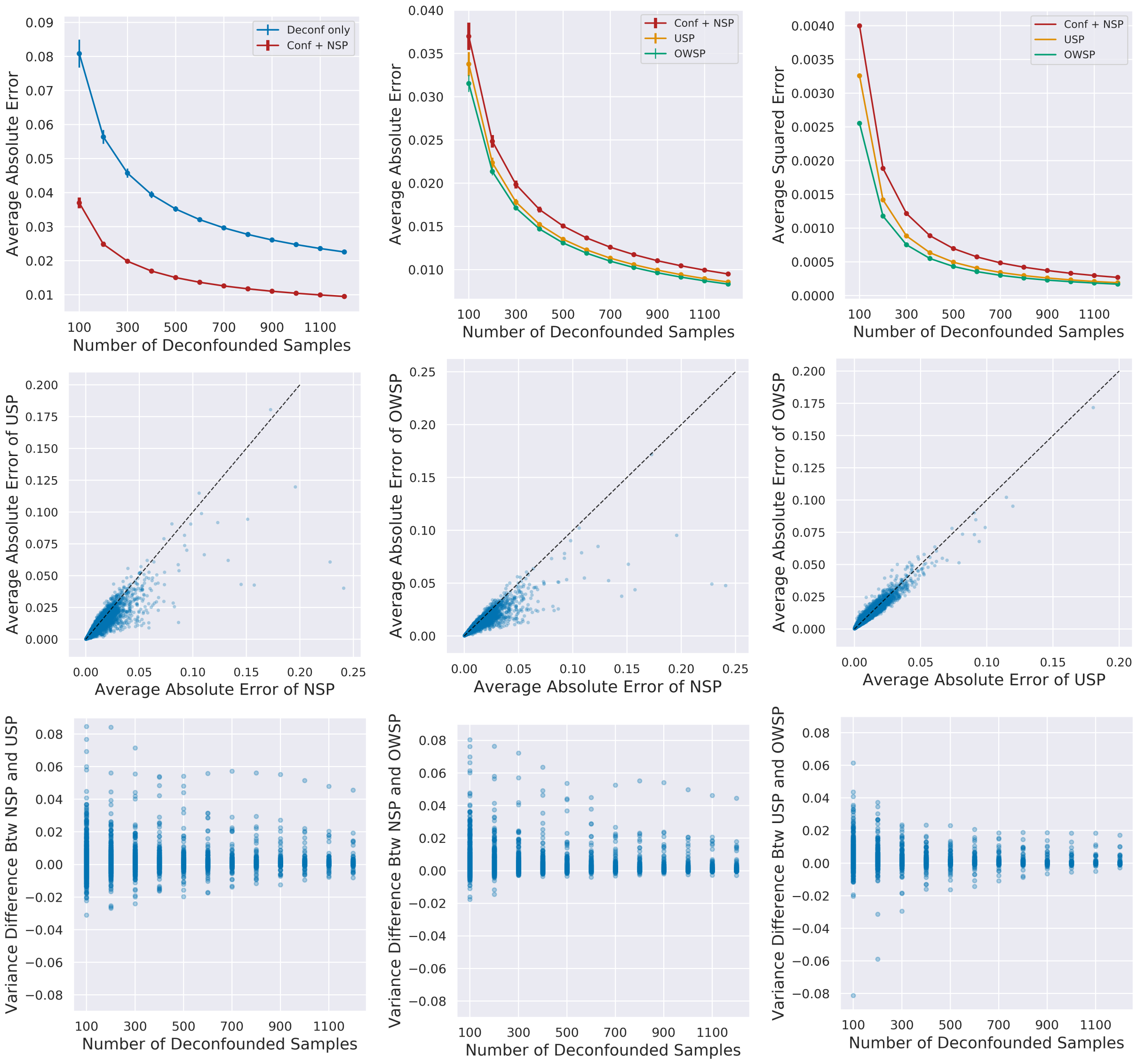


Figure 3.2: Performance of the four sampling policies over 13,000 distributions $\mathbb{P}_{Y,T,Z}$, assuming infinite confounded data. Top row (left and middle): averaged absolute error over all 13,000 distributions for varying numbers of deconfounded samples. Top row (right): averaged squared error over all 13,000 distributions. Middle row: error comparison (each point is a single distribution averaged over 100 replications) for 1,200 deconfounded samples. Bottom row: variance comparison (each point corresponds to one of the 13,000 distributions and the variance is calculated over the 100 replications) between selected sample selection policies. Bottom row: each dot corresponds to the *difference* between the variance of a pair of selected methods under one instance and a fixed number of deconfounded samples. Bottom left: a positive y-axis value implies that USP yields a smaller variance than NSP. Bottom middle: a positive y-axis value implies that OWSP yields a smaller variance than NSP. Bottom right: a positive y-axis value represents that OWSP yields a smaller variance than USP.

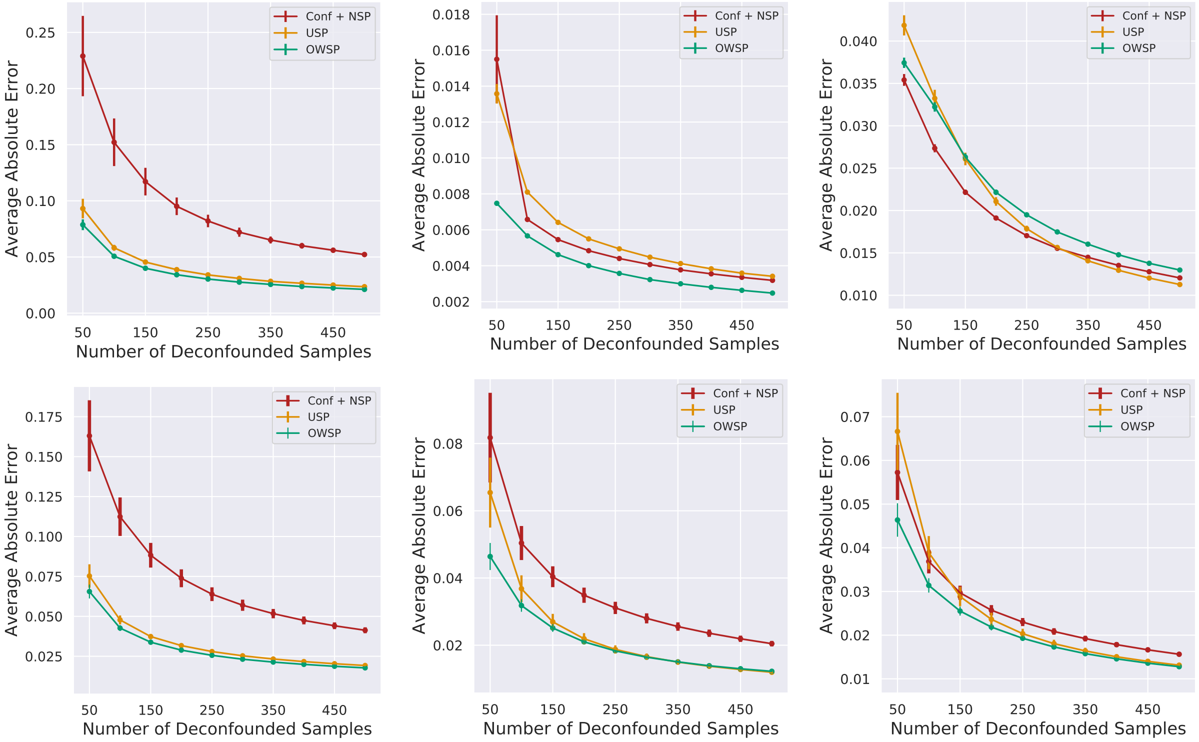


Figure 3.3: Comparison of selection policies for adversarially chosen instances. Top row left: $\mathbf{a} = (0.9, 0.02, 0.01, 0.07)$ and $\mathbf{q} = (0.9, 0.7, 0.01, 0.3)$, where NSP performs the worst. Top row middle: $\mathbf{a} = (0.79, 0.01, 0.02, 0.18)$ and $\mathbf{q} = (0.5, 0.01, 0.05, 0.5)$, where USP performs the worst. Top row right: $\mathbf{a} = (0.5, 0.01, 0.19, 0.3)$ and $\mathbf{q} = (0.05, 0.5, 0.055, 0.4)$, where OWSP performs the worst. Bottom row: generated with the same $\mathbf{a}$'s but averaged over 500 $\mathbf{q}$'s drawn uniformly from $[0, 1]^4$.

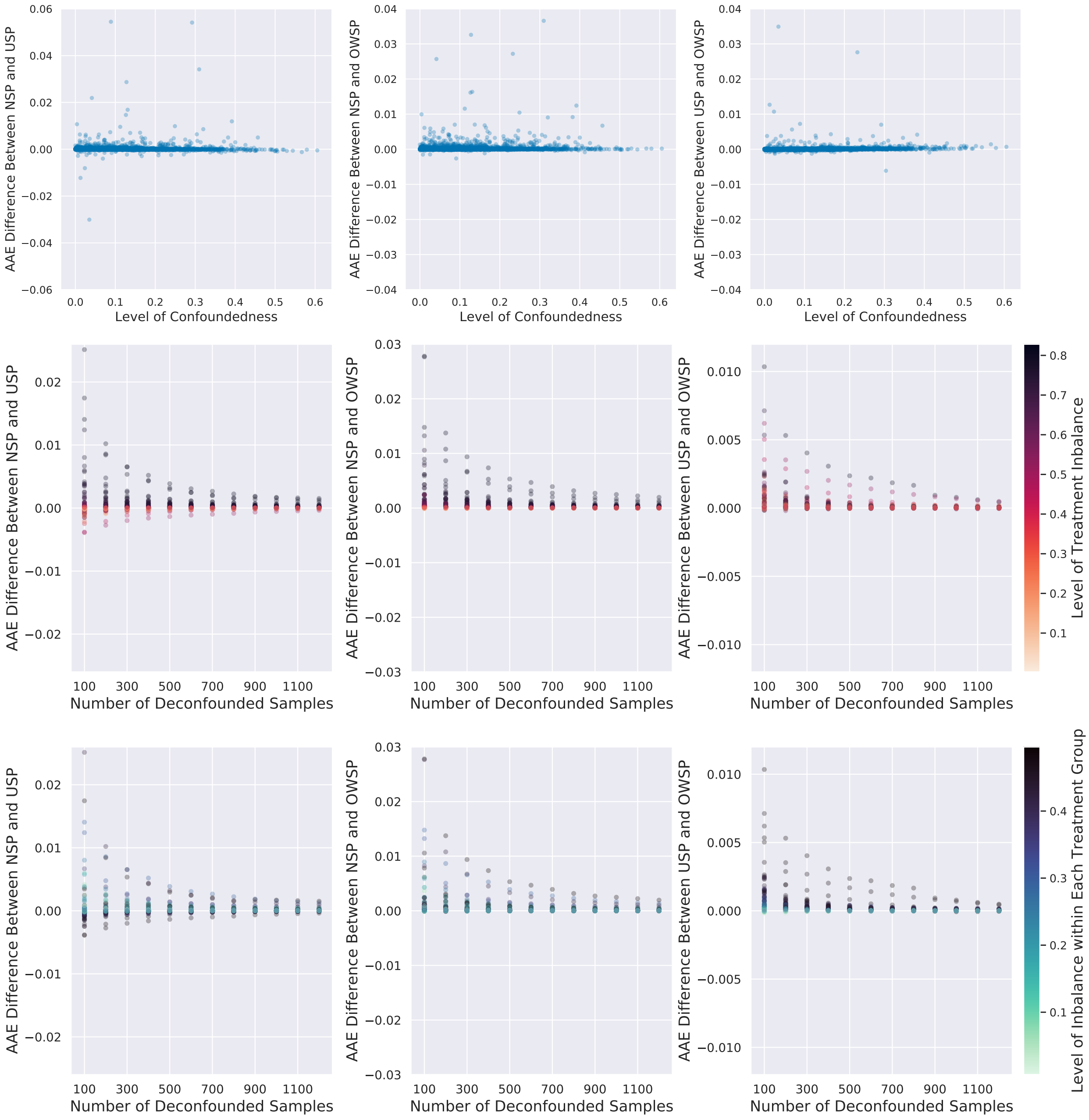


Figure 3.4: Performance Insights of the four sampling policies over a different set of 13,000 distributions $\mathbb{P}_{Y,T,Z}$, assuming infinite confounded data. The y-axis of all figures is the *average absolute error* (AAE) difference between a pair of selected methods. Each instance is averaged over 100 replications. Top row: contains 13,000 dots, each representing an instance. The number of confounded data is fixed at 1200. The x-axis is the level of confoundedness of an instance. Middle and Bottom rows: at each level of deconfounded samples, each figure contains 130 dot, each representing one confounded distribution **a**, averaged over 100 conditional distributions **q**. The x-axis is the number of deconfounded samples measured in steps of 100. Middle row: the color map corresponds to the level of treatment inbalance of an instance. Bottom row: the color map corresponds to the level of (maximum) outcome inbalance within each treatment group of an instance.

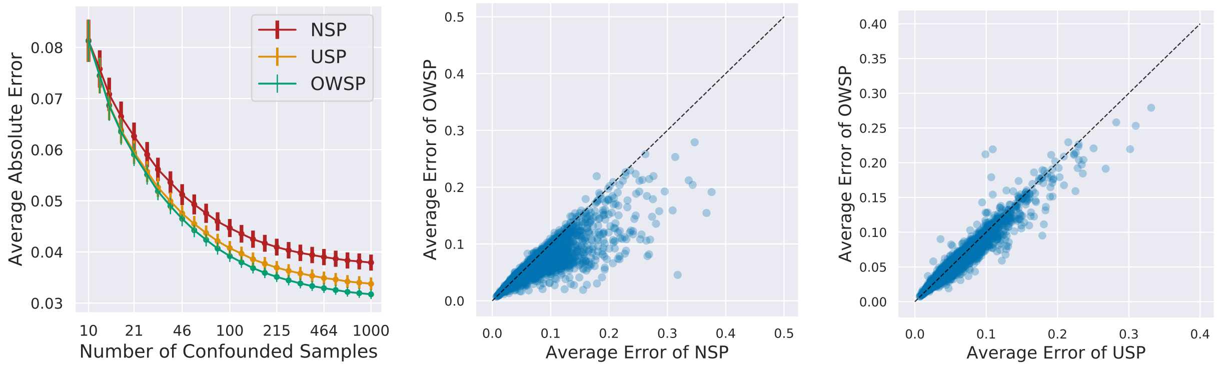


Figure 3.5: Experiment on finite confounded data over 13,000 distributions $P_{Y,T,Z}$, each averaged over 100 replications. The number of deconfounded samples is fixed at 100. Left: averaged over the 13,000 distributions. Middle and Right: error comparison at 681 confounded samples.

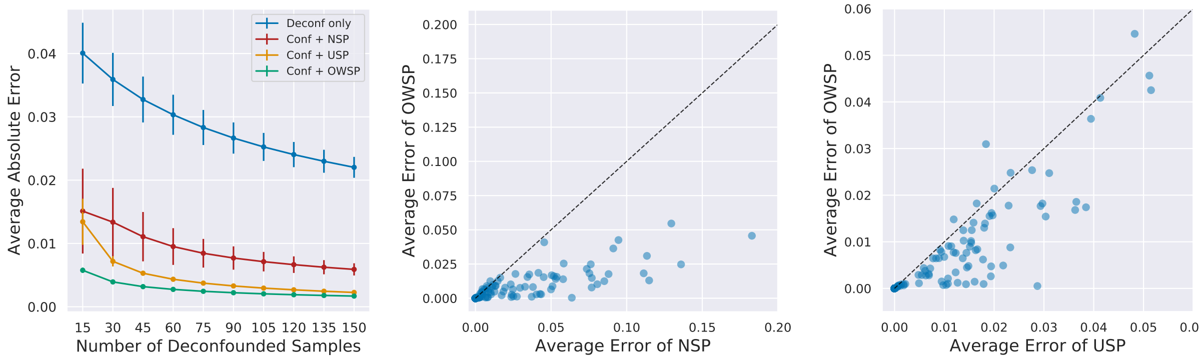


Figure 3.6: Performance of the four sampling policies on the COSMIC dataset assuming infinite confounded data. 275 unique (cancer, mutation, mutation) combinations were extracted. Left: averaged over 275 instances, and each averaged over 10,000 replications. Middle and Right: error comparison at 45 deconfounded samples.

Chapter 4

Toward a Liquid Biopsy: Greedy Approximation Algorithms for Active Sequential Hypothesis Testing

4.1 Introduction

Among the most important open problems in cancer research today is the development of an effective approach for the *detection* of cancer, particularly at its *earliest* stages. Indeed for quite some time now, there has been vast, uncontroverted evidence that early detection substantially enhances the possibility of successful treatment (Etzioni et al. 2003, Cuzick et al. 2014, Jerant et al. 2000). As a few examples, the five-year survival rates after early diagnosis (and treatment) of breast, ovarian, and lung cancers are 90%, 90%, and 70%, respectively, compared to 15%, 5%, and 10% for patients diagnosed at the latest stages (Siegel et al. 2015, Jemal et al. 2010, Ferguson et al. 2000). In short, early detection is a silver bullet.

Unfortunately, although monitoring certain “warning signs” occasionally yields early diagnoses, cancer screening is in general notoriously difficult, and existing approaches fall short. Modern cancer diagnoses are for the most part made via *biopsies*, i.e., the surgical removal of tissue for testing, and while biopsies are extremely accurate (with respect to identifying cancer in the removed tissue itself), they are too invasive and expensive to be used as a general screening procedure.[1] But even beyond issues like cost and inconvenience to the patient, the use of biopsies for *early* cancer screening is in fact fundamentally impossible for several cancer types, such as lung and pancreatic cancers, which almost never show symptoms until after cancer cells have

[1] Less invasive screening tools do exist, but by and large none has achieved the requisite accuracy to be adopted by the medical community for general screening.

metastasized[2] (Miller et al. 1981, Paez et al. 2004, O'Rourke and Edwards 2000).

Because of these difficulties, there has always existed a dream within the medical community of developing a *liquid biopsy*, i.e., a blood test for cancer. This test would naturally be minimally invasive, and ideally would have the same accuracy as a traditional biopsy. Most importantly, the liquid biopsy would detect cancers at their earliest stages. What is particularly exciting today is that liquid biopsies are no longer just a pipe dream – these tests have been under rapid development over the last few years, largely fueled by advances in technology for collecting data (next-generation DNA sequencing, in particular), and increasing computational and algorithmic power for analyzing this new data. Development of these tests is being undertaken by major academic research labs (Bettegowda et al. 2014, Manterola et al. 2014, Best et al. 2015, Kim et al. 2016, Banerjee et al. 2016, Razavi et al. 2017, Chan et al. 2017, Cohen et al. 2018, Liu et al. 2020) along with a handful of biotechology startups (e.g., Grail, Guardant, Freenome).

4.1.1 The Genomic Approach and the Value of Adaptivity

Ultimately, this paper addresses a set of *active learning* problems that occur in the development of liquid biopsies. To understand how such problems arise, it may be useful to review, at a high level, the underlying biology here (this subsection can be skipped without loss of continuity). Starting with a basic fact: cancer is caused by mutations in DNA, meaning the DNA within every tumor cell has a *set* of mutations that identifies the cell as tumorous, along with its location in the body (and thus the type of cancer).[3] These mutations are the "signals" that *genomic* liquid biopsies are designed to detect. The reason that these signals are detectable from blood is due to *cell-free DNA*—the DNA of any dying cell is occasionally released into the bloodstream (rather than being destroyed), and so an individual's blood at any moment contains free floating DNA that we can view as having been randomly "sampled" (in a probabilistic sense) from throughout the body.

Thus comes the main idea: if an individual has a tumor, some portion of their cell-free DNA will contain mutations which signal the existence of that tumor. So performing the liquid biopsy simply involves extracting cell-free DNA (a relatively easy task), and sequencing it in search of these mutations. There is no purely biological reason why this approach should fail. Instead, the constraint that we face today is the *cost*—human DNA consists of three billion *addresses*, but the cost of DNA sequencing means that any reasonably-priced test can only include approximately

[2] *Metastasis* refers to the formation of a new cancer "colony" at a separate location in the body. The occurrence of this process is generally used to define the line between early and late stage cancers, as once a primary tumor metastasizes, successful treatment using established therapeutics becomes nearly hopeless.

[3] Identifying the specific mutations which *cause* a particular cancer to form is still an open problem. Fortunately for the purposes of a liquid biopsy, correlation suffices.

10^4 of those addresses.[4] So the challenge is to design a liquid biopsy using just a *panel* of 10^4 pre-identified addresses.

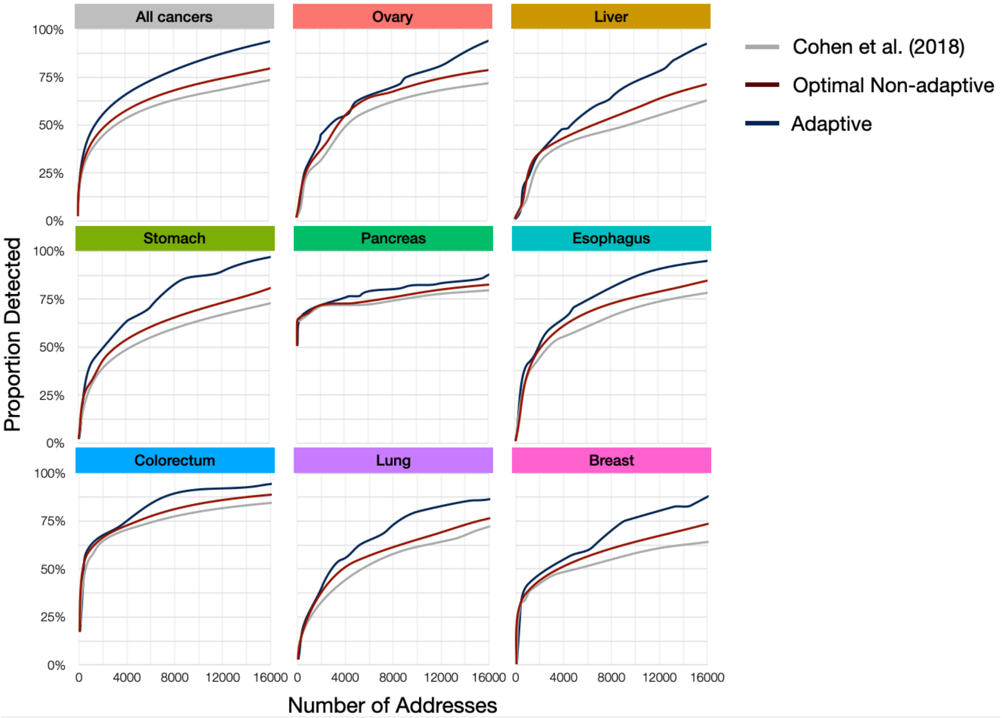


Figure 4.1: Comparison of (non-adaptive) genomic panels from Cohen et al. (2018) with *optimal* non-adaptive panels, and adaptive panels constructed using our greedy adaptive algorithm. For each approach, detection rate (on the COSMIC dataset) is plotted as a function of panel size. Results are reported for eight cancer types, in combination and individually. Figure format adapted from Cohen et al. (2018).

To date, one of the two most successful prototypes of a liquid biopsy is from Cohen et al. (2018).[5] To select the panel of DNA addresses for their liquid biopsy, they used the publicly-available dataset—*Catalogue of Somatic Mutations in Cancer* (COSMIC, Tate et al. 2019) which contains complete DNA sequences from thousands of tumor cells, and this allows one to “simulate” the accuracy of different combinations of addresses subject to any budget constraint and select the

[4] The back-of-the-envelope calculation works as follows: modern DNA sequencing costs approximately $\$10^{-6}$ (USD) per address. But because tumorous DNA only makes up a tiny proportion (about one in ten-thousand) of a cancer patient’s cell-free DNA, each address used in a liquid biopsy needs to be sequenced 10^4 times (to avoid false negatives). Finally, a screening test should cost at most $\$10^2$, so $\$10^2$ / ($\$10^{-2}$ per address) = 10^4 addresses. In the adaptive problem that we will introduce later, we further relax the assumption that each address needs to be sequenced 10^4 times. Instead, the number of repetition will be determined by our algorithms.

[5] The other is Liu et al. (2020).

most accurate one.[6] Cohen et al. (2018) did exactly this analysis, which is reproduced here as the grey-colored curves in Fig. 4.1. Each of those curves shows how the proportion of detected cancer patients increases with the number of addresses, for eight different cancer types.

Now the problem of identifying the subset of addresses, subject to cardinality constraint, that maximizes the number of cancer patients detected (within the COSMIC dataset) is a well-defined optimization problem (a maximum coverage problem, in fact), and while this application can become quite large in practice (technically, each of the three billion DNA addresses is a "set," and there are tens of thousands of samples in COSMIC to "cover"), this particular instance can be solved to optimality with an off-the-shelf integer programming solver. That is precisely what we have done and represented as the red curves in Fig. 4.1, and we can see that the red curves indeed dominate the grey curves of Cohen et al. (2018).

After a perhaps long-winded introduction, it is at *this* point that the problem we seek to address has finally appeared. Reviewing Fig. 4.1 again, an unfortunate observation is that even the optimal panels (the red curves) are insufficient for a practical liquid biopsy—in visual terms, the curves do not reach far enough into the top-left area (representing low cost and high accuracy). Now advances in DNA sequencing technology may eventually solve this problem (by further reducing sequencing costs), but the purpose of this paper is to study a more immediate solution: *adaptive testing*. We use the term "adaptive" to mean that the test for an individual can operate over multiple rounds, with the choice of addresses in each round being made using the results of prior rounds (the tests used by Cohen et al. (2018), along with our "optimal" panels, were non-adaptive). The problem of identifying the optimal adaptive test can similarly be formalized (as we will do), though that problem almost certainly does not admit a solution via computational brute force. Instead, we will analyze fast approximation algorithms, whose practical value is partly demonstrated by the blue curves in Fig. 4.1.

4.1.2 The Problem and Our Contributions

At this point, it is worth abstracting away the application, because the natural model for this is a well-studied one. Consider the problem of learning the *true* hypothesis from among a (potentially large) set of candidate hypotheses H. Assume that the learner is given a (potentially large) set of actions A, and knows the distribution of the noisy outcome of each action, under each potential hypothesis. In the context of liquid biopsies, the candidate hypotheses are different types of cancers, and the actions correspond to sequencing individual DNA addresses (actually combinations of

[6] Readers with experience in machine learning might interpret this entire procedure as training a classifier, with COSMIC as the training set, and be concerned about over-fitting and generalization error. While we view this as orthogonal to the problem we seek to address, it is worth noting that Cohen et al. (2018) actually observed *higher* detection rates (i.e., accuracy) in practice than those predicted by COSMIC.

	Noise	**Approx. Ratio**	**Objective**	**Adaptivity Type**
Kosaraju et al. (1999)	No	Yes	Both	Full
Chakaravarthy et al. (2009)	No	Yes	Both	Full
Nowak (2009)	Yes	No	Worst-case	Full
Naghshvar and Javidi (2013)	Yes	No	Average	Both
Im et al. (2016)	No	Yes	Both	Partial
Jia et al. (2019)	Semi*	No	Both	Both
This Work	Yes	Yes	Both	Both

Table 4.1: Summary of related work. **Semi* refers to a restrictive special case. Approx. stands for approximation.

addresses, as we will discuss later on). The learner incurs a fixed cost each time an action is selected, and seeks to identify the true hypothesis with sufficient confidence, at minimum total cost. Finally, and most importantly, the learner is allowed to select actions *adaptively*.

This well-studied problem is referred to as *active sequential hypothesis testing*, and as we will describe momentarily, there exists a broad set of results that tightly characterizes the optimal achievable cost under various notions of adaptivity. Unfortunately, the corresponding optimal policies are typically only characterized as the optimal policy to a Markov decision process (MDP)—thus, they remain computationally hard to compute when one requires a policy in practice. This deficiency becomes particularly apparent in modern applications where both the set of hypotheses and set of actions may be large: our own application has tens of hypotheses and *billions* of tests at full scale. Thus motivated, *our primary contribution is the first approximation algorithms for ASHT*.

We study ASHT under two notions of adaptivity: *partial* and *full*, where partial adaptivity requires the sequence of actions to be decided upfront (with adaptively chosen stopping time), and full adaptivity allows the choice of action to depend on previous outcomes. For both problems, we propose *greedy* algorithms that run in $O(|A||H|)$ time, and prove that their expected costs are upper bounded by a non-trivial multiplicative factor of the corresponding optimal costs. Most notably, these approximation guarantees are *independent* of $|A|$ (contrast this with the trivially-achievable guarantee of $O(|A|)$) and *logarithmic* in $|H|$ (the optimal cost itself is often $\Omega(|H|)$).

Our theoretical results rely on drawing connections to two existing problems: *submodular function ranking* (SFR, see Azar and Gamzu 2011) and the *optimal decision tree* (ODT) problem (Laurent and Rivest 1976). These connections allow us to tackle what is arguably the primary challenge in achieving approximation results for ASHT, which is its inherent *combinatorial* nature. We will argue that existing heuristics from statistical learning fail precisely because they disregard this combinatorial difficulty—indeed, they largely amount to solving the completely *non-adaptive* version of the problem. At the same time, existing results for SFR and ODT fail to account for *noise* in a manner that would map directly to ASHT—this extension is among our contributions.

Finally, we performed a set of large-scale experiments, including the ones that were built on the same setup of Cohen et al. (2018). These experiments demonstrate that, in both the partial and fully adaptive settings, our greedy algorithms (a) scale to the size of real-world problems, and (b) outperform existing benchmarks for ASHT in practice. In the setting of liquid biopsies, our results suggest that adaptive testing is sufficient to the achieve the remaining accuracy needed to bring about a practical cancer screening test.

4.1.3 Related Work

Our work is closely related to three streams of research. Table 4.1 highlights the key differences between our contributions and those of the most relevant previous works.

Hypothesis Testing and Asymptotic Performance In the classical binary sequential hypothesis testing problem, a decision maker is provided with one action whose outcome is stochastic (Wald 1945, Armitage 1950, Lorden 1977), and the goal is to use the minimum expected number of samples to identify the true hypothesis subject to some given error probability. The ASHT problem, first studied in Chernoff (1959), generalizes this problem to multiple actions. Most related to our work is Naghshvar and Javidi (2013), who formulated a similar problem as an MDP. We will postpone describing and contrasting their work until the experiments section.

Active Learning and Sample Complexity In active learning, the learner is given access to a pool of unlabeled samples (cheaply obtainable) and is allowed to request the label of any sample (expensive) from that pool. The goal is to learn an accurate classifier while requesting as few labels as possible. Some nice surveys include Hanneke et al. (2014) and Settles (2009). Our model extends the classical discrete active learning model Dasgupta (2005) in which outcomes are noiseless (deterministic) for any pair of hypotheses and unlabeled sample. When outcomes are noisy, the majority of provable guarantees are provided via sample complexity using minimax analysis techniques. Castro and Nowak (2007) showed tight minimax classification error rates for a broad class of distributions. Other sample complexity results on noisy active learning include Wang and Singh (2016), Nowak (2009), Balcan et al. (2006), Awasthi et al. (2017), Hanneke and Yang (2015).

Approximation Algorithms for Decision Trees Nearly all optimal approximation algorithms for minimizing cover time are known in the noiseless setting (Kosaraju et al. 1999, Adler and

Heeringa 2008, Arkin et al. 1998). When the outcome is stochastic, Golovin and Krause (2011) proposed a framework for analyzing algorithms in the active learning setting under the *adaptive submodularity* assumption, with the goal of maximizing the information gained with a prescribed budget. However, their assumption does not hold for many natural setups including ASHT. Chen et al. (2015) considered a variant using ideas from the submodular max-coverage problem without the adaptive submodularity assumption, and provided a constant factor approximation to the problem using ideas from the submodular max-coverage problem. Other works based on submodular function covering include Navidi et al. (2020), Guillory and Bilmes (2011), Krause et al. (2008). Jia et al. (2019) provided approximation ratios under the constraint that the algorithm may only terminate when it is completely confident about the outcome.

4.2 Model

In this section, we formally introduce the active sequential hypothesis testing problem; the mapping of this generic problem to the liquid biopsy application is described in detail in § 4.6. Let H be a finite set of *hypotheses*, among which exactly one is the (unknown) *true* hypothesis that we seek to identify. We denote this true hypothesis h^*. In this paper, we study the *Bayesian* setting, wherein h^* is drawn from a known prior distribution π over the entire hypothesis set H.

Let A be the set of available *actions*, and let $\mathcal{D}$ be a given family of parametrized distributions that encode the *outcome distributions*, i.e., the distributions from which the outcome is drawn when an action is chosen. For ease of exposition, we parameterize the family of distributions $\mathcal{D}$ by $\Theta \subseteq \mathbb{R}$, i.e., $\mathcal{D} = \{D_\theta\}_{\theta\in\Theta}$. Thus, selecting, or "playing", an action yields a random *outcome* drawn independently from a distribution within the given family $\mathcal{D} = \{D_\theta\}_{\theta\in\Theta}$. In addition, we are given a function μ that maps each pair of action and hypothesis to some $\theta \in \Theta$, i.e., $\mu : H \times A \longrightarrow \Theta$, such that if $h \in H$ is the true underlying hypothesis and we select action $a \in A$ to play, then we observe a random outcome that is drawn independently from distribution $D_{\mu(h,a)}$. Note that in this *noisy* setting, an action can (and often should) be played multiple times (in the same way that a DNA address should be sequenced multiple times since each strand of cell-free DNA is effectively sampled randomly from throughout the body).

An *instance* of the active sequential hypothesis testing problem is then fully specified by a tuple: $(H, A, \pi, \mu, \mathcal{D})$. The goal is to sequentially select actions to identify the true hypothesis with "sufficiently high" confidence, at minimal expected cost, where cost is measured as the number of actions, and the expectation is with respect to the Bayesian prior and the random outcomes. The notion of *sufficiently high* confidence is encoded by a parameter $\delta \in (0, 1)$, and requires that under any true hypothesis $h \in H$, the probability of erroneously identifying a different hypothesis is at most δ. An algorithm is said to have achieved **δ-PAC-error** if it identifies the true hypothesis with this notion of sufficiently high confidence.

In this paper, we focus on two important families of the outcome distributions, D_θ's: the Bernoulli distribution $\mathrm{Ber}(\theta)$, and the Gaussian distribution $N(\theta, \sigma^2)$. In the latter, the variance σ^2 is a known constant (with respect to θ).[7] By re-scaling, without loss of generality we may assume $\sigma^2 = 1$. To state our guarantees, we require two additional assumptions. The first assumption is needed for relating the sub-gaussian norm to the KL-divergence, in the partially adaptive setting. It ensures that the parameterization Θ is a meaningful one, in the sense that if two parameters $\theta, \theta' \in \Theta$ are far apart, then the corresponding distributions D_θ and $D_{\theta'}$ are also "far" apart, (i.e., their KL divergence is "far" apart). Note that Assumption 2 is satisfied for the Bernoulli distribution $\mathrm{Ber}(\theta)$ when $\theta \in [\theta_{\min}, \theta_{\max}]$ for some constants $0 < \theta_{\min} < \theta_{\max} < 1$, and for the Gaussian distribution $N(\theta, 1)$ where θ lies in some bounded subset of $\mathbb{R}$.

Assumption 2. *There exist two constants $C_1, C_2 > 0$ such that for any $\theta, \theta' \in \Theta$, we have*

$$C_1 \cdot \mathrm{KL}(D_\theta, D_{\theta'}) \leq (\theta - \theta')^2 \leq C_2 \cdot \mathrm{KL}(D_\theta, D_{\theta'}),$$

where $\mathrm{KL}(\cdot, \cdot)$ is the Kullback-Leibler divergence.

Our second major assumption simply ensures the existence of a valid algorithm by ensuring that every hypothesis is distinguishable via some action:

Assumption 3 (Validity)**.** *For any pair of distinct hypotheses $g, h \in H$, there exists an action $a \in A$ with $\mu(g, a) \neq \mu(h, a)$.*

Note that in Assumption 3, we do not preclude the possibility that for a given action a, there exist (potentially many) pairs of hypotheses g and h such that the outcome distributions are the same, i.e., $\mu(g, a) = \mu(h, a)$. In fact, eliminating such possibilities would effectively wash out any meaningful combinatorial dimension to this problem. On the other hand, any approximation guarantee should be parameterized by some notion of *separation* (when it exists). For any two hypotheses $g, h \in H$ and any action $a \in A$, we define the distance between these two hypotheses under action a as $d(g, h; a) := \mathrm{KL}\left(D_{\mu(g,a)}, D_{\mu(h,a)}\right)$. Let $s > 0$ be some positive constant. The following is the notion of separation, *s-separability*, that we use throughout the paper:

Definition 4 (s-separated instance)**.** *An ASHT instance is said to be s-**separated**, if for any $a \in A$ and $g, h \in H$, $d(g, h; a)$ is either 0 or at least s.*

Note that in real-world applications, the parameters s could be arbitrarily small, and we introduce the notation of s-separability for the sake of proofs. We will show in § 4.6 how our algorithms can easily be modified to handle small s values. In this paper, we will study two classes of algorithms that differ in the extent to which adaptivity is allowed.

[7] Sub-Gaussianity with similar control over the sub-Gaussian norm would suffice.

Definition 5. *A **fully adaptive** algorithm is a decision tree,*[8] *each of whose interior nodes is labeled with some action, and each of whose edges corresponds to an outcome. Each leaf is labeled with a hypothesis, corresponding to the output when the algorithm terminates.*

Definition 6. *A **partially adaptive** algorithm* (σ, T) *is specified by a fixed sequence of actions* $\sigma = (\sigma_1, \sigma_2, ...)$*, with each* $\sigma_i \in A$*, and a stopping time* T*, such that the event* $\{T = t\}$ *is independent of the outcomes of actions* $\sigma_{t+1}, \sigma_{t+2}, ...,$ *under any true hypothesis* $h^* \in H$ *and for any* $t \geq 1$*. (At the stopping time, the choice of which hypothesis to identify is trivial in our Bayesian setting—it is simply the one with the highest "posterior" probability.)*

Note that a partially adaptive algorithm can be viewed as a special type of fully adaptive algorithm: it is a decision tree with the additional restriction that the actions at each depth are the same. Therefore, a fully adaptive algorithm may be far cheaper than any partially adaptive algorithm. However, there are many scenarios (e.g., content recommendation and web search (Azar et al. 2009)) where it is desirable to fix the sequence of actions in advance. Furthermore, in many problems the theoretical analysis of partially adaptive algorithms turns out to be challenging (e.g., Kamath and Tzamos 2019, Chawla et al. 2019).

Thus, given an ASHT instance, there are two problems that we will consider, depending on whether the algorithms are partially or fully adaptive. In both cases, our goal is to design fast approximation algorithms—ones that are computable in polynomial[9] time and that are guaranteed to incur expected costs at most within a multiplicative factor of the optimum. In the coming sections, we will describe our algorithms and approximation guarantees. Before moving on to this, it is worth noting that our problem setup is extremely generic and captures a number of well-known problems related to decision-making for learning including best-arm identification for multi-armed bandits (Bubeck et al. 2009, Even-Dar et al. 2002, Mannor and Tsitsiklis 2004), group testing (Du et al. 2000), and causal inference (Gan et al. 2020), just to name a few.

4.3 Our Approximation Guarantees

In this section, we will state our approximation guarantees. We will define the corresponding greedy algorithms in the next two sections. Let $\mathrm{OPT}^{\mathrm{PA}}_{\delta}$ and $\mathrm{OPT}^{\mathrm{FA}}_{\delta}$ denote the minimal expected cost of any partially adaptive and fully adaptive algorithm that achieves δ-PAC-error, respectively. Theorem 11 summarizes the approximation guarantee for our greedy partially adaptive algorithm:

[8] By approximating D_θ's with discrete distributions, we may assume each node has a finite number of children.

[9] Throughout this paper, *polynomial time* refers to polynomial in $\left(|H|, |A|, s^{-1}, \delta^{-1}\right)$

Theorem 11. *Given an s-separated instance and any $\delta \in (0, 1/2)$, there exists a polynomial-time partially adaptive algorithm that achieves δ-PAC-error with expected cost*

$$O\left(s^{-1}\left(1 + \log_{1/\delta}|H|\right)\log\left(s^{-1}|H|\log\delta^{-1}\right)\right)\mathrm{OPT}_{\delta}^{\mathrm{PA}}.$$

To help parse this result, if δ is on the order of $|H|^{-c}$ for some constant c, then the approximation factor becomes $s^{-1}(\log s^{-1} + \log|H| + c\log\log|H|)$. Theorem 12 summarizes the approximation guarantee for our greedy fully adaptive algorithm:

Theorem 12. *Given an s-separated instance and any $\delta \in (0, 1/2)$, there exists a polynomial-time fully adaptive algorithm that achieves δ-PAC-error with expected cost*

$$O\left(s^{-1}\log\left(|H|\delta^{-1}\right)\log|H|\right)\mathrm{OPT}_{\delta}^{\mathrm{FA}}.$$

A few observations might clarify the significance of these approximation guarantees:

1. Dependence on action space: Both guarantees are independent of the number of actions $|A|$. This is extremely important since, as described in the Introduction, there exist many applications where the the action space is massive. Moreover, since an approximation factor of $O(|A|)$ is always trivially achievable (by cycling through the actions), instances where $|A|$ is large are arguably the most interesting problems.

2. Dependence on $|H|$, δ and s: For fixed s and δ, these are the first polylog-approximations for both partially and fully adaptive versions. Further, for the partially adaptive version, the dependence of the approximation factor on δ is $O(\log\log\delta^{-1})$ when δ^{-1} is polynomial in $|H|$, improving upon the naive dependence $O(\log\delta^{-1})$. This is crucial since δ is often needed to be tiny in practice.

3. Greedy runtime: While we have only stated in our formal results that our approximation algorithms can be computed in poly$(|A|, |H|)$ time, the actual time is more attractive: $O(|A||H|)$ for selecting each action. In contrast, the heuristic that we will compare against in the experiments requires solving multiple $\Omega(|A||H|^2)$-sized linear programs.

Despite their similar appearances, Theorems 11 and 12 rely on fundamentally different algorithmic techniques and thus require different analyses. In § 4.4, we propose an algorithm inspired by the *submodular function ranking* problem, which greedily chooses a sequence of actions according to a carefully chosen "greedy score." We then sketch the proof of Theorem 11. In § 4.5, we introduce our fully adaptive algorithm and sketch the proof of Theorem 12.

Finally, by proving a structural lemma (in § 4.11), we extend the above results to a special case of the **total-error** version (i.e., averaging the error over the prior π) where the prior distribution is uniform. With *δ-total-error* formally defined in § 4.11:

Theorem 13. *Given an s-separated instance with uniform prior π and any $\delta \in (0, \frac{1}{4})$, for both the partially and fully adaptive versions, there exist polynomial-time δ-total-error algorithms with expected cost $O\left(s^{-1}\left(1 + |H|\delta^2\right) \log\left(|H|\delta^{-1}\right) \log |H|\right)$ times the optimum.*

4.4 Partially Adaptive Algorithm

This section describes our algorithm and guarantee for the partially adaptive problem. We first review necessary background from a related problem, and then state our algorithm (Algorithm 8). Finally, we sketch the proof of the following more general version of Theorem 11 (complete proof in § 4.9):

Proposition 9. *Let $\delta \in (0, \frac{1}{4}]$ and consider finding the optimal δ-PAC error algorithm. Given any boosting intensity $\alpha \geq 1$ and coverage saturation threshold $B \in (0, \frac{1}{2}\log \delta^{-1}]$, $\mathrm{RnB}(B, \alpha)$ (as defined in Algorithm 8) produces a partially adaptive algorithm with error $|H| \exp\left(-\Omega\left(\alpha B\right)\right)$ and expected cost $O\left(\frac{\alpha}{s} \log \frac{|H|B}{s}\right) \mathrm{OPT}_{\delta}^{\mathrm{PA}}$.*

By setting $\alpha = 1 + \log_{\delta^{-1}} |H|$ and $B = \frac{1}{2}\log \delta^{-1}$, we immediately obtain Theorem 11.

4.4.1 Background: Submodular Function Ranking

In the SFR problem, we are given a ground set U of N *elements*, a family $\mathcal{F}$ of non-decreasing submodular functions $f : 2^U \longrightarrow [0, 1]$ with $f(U)$ equaling 1 for every $f \in \mathcal{F}$, and a weight function $w : \mathcal{F} \longrightarrow \mathbb{R}^+$. For any permutation $\sigma = (u_1, ..., u_N)$ of U, the *cover time* of f is defined as $\mathrm{CT}(f, \sigma) = \min\{t : f(\{u_1, ..., u_t\}) = 1\}$. The goal is to find a permutation σ of U with minimal weighted *cover time*, $\sum_{f \in \mathcal{F}} w(f) \cdot \mathrm{CT}(f, \sigma)$.

A greedy algorithm was proposed in Azar and Gamzu (2011), and we will use this algorithm as an important subroutine in our algorithm. This greedy algorithm constructs a sequence iteratively. At each iteration, we say a function is *covered* if its value on the set of the elements selected so far is 1, and the function is *uncovered* otherwise. The sequence is initialized to be empty. At each iteration, let S denote the set of elements selected so far. The algorithm selects the element u with the maximal *coverage*, defined as

$$\mathrm{Cov}(u; S) := \sum_{f \in \mathcal{F} : f(S) < 1} w(f) \cdot \frac{f(S \cup \{u\}) - f(S)}{1 - f(S)}.$$

Loosely, the algorithm chooses the element that maximizes the weighted sum of additional immediate proportional coverage. The following approximation guarantee for this algorithm is known to be the best possible among all polynomial-time algorithms:

Theorem 14 (Im et al. 2016). *For any SFR instance, the greedy algorithm described above returns a sequence whose cost is $O(\log \varepsilon^{-1})$ times the optimum, where*

$$\varepsilon := \min \left\{ f(S \cup \{u\}) - f(S) > 0 \,:\, S \in 2^U, u \in U, f \in \mathcal{F} \right\}.$$

Challenge: To motivate our algorithm, consider first the following simple idea: "boost" each action, and hence reduce the problem to a deterministic problem P_{det}. Then show that the existing technique (submodular function ranking for partially adaptive and greedy analysis for ODT for fully-adaptive) returns a policy with cost $O(\log |H|)$ times the no-noise optimum, and finally show that this no-noise policy can be converted to a noisy version by losing anther factor of $O(s^{-1} \log(\delta^{-1}|H|))$. This analysis was in fact our first attempt. However, there are at least two issues that one runs into along this path:

1. This analysis only compares the policy's cost with the no-noise optimum, but our focus is the δ-noise optimum. In particular, the simpler analysis implicitly assumes that the δ-noise optimum is at least $\Omega(s^{-1} \log(\delta^{-1}|H|))$ times the no-noise optimum, which is not necessarily true. Moreover, it is challenging to analyze the gap between the no-noise optimum and the δ-noise optimum.

2. The guarantee that results from this simple analysis is *weaker* than ours in terms of δ: it yields a factor of $\log(1/\delta)$, as opposed to the $\log\log(1/\delta)$ in our analysis. This distinction is nontrivial, particularly in applications where the error is required to be exponentially small in $|H|$.

4.4.2 Rank and Boost (RnB)

Our RnB algorithm (Algorithm 8) circumvents the issues above by drawing a connection between ASHT and SFR. First, we observe that although an action is allowed to be selected multiple times, we may assume each action is selected for at most $M = M(\delta, s, |H|) = O(s^{-1}|H|^2 \log(|H|/\delta))$ times. In fact,

Observation 1. *Let $\widetilde{A}$ be the (multi)-set obtained by creating M copies of each $a \in A$. Then there exists a sequence σ of $|\widetilde{A}|$ actions, such that h^*, the true hypothesis, has the highest posterior with probability $1 - \delta$ after performing all actions in σ.*

Thus, given $\widetilde{A}$, we define $f_h^B : 2^{\widetilde{A}} \longrightarrow [0,1]$ for any coverage saturation level $B > 0$ and $h \in H$ as $f_h^B(S) = \frac{1}{|H|-1} \sum_{g \in H \setminus \{h\}} \min\{1, \frac{1}{B} \sum_{a \in S} d(g, h; a)\}$. One can verify that f_h^B is monotone and submodular. Our algorithm computes a nearly optimal sequence of actions using the greedy algorithm for SFR, and creates a number of copies for each of them. Then we assign a *timestamp* to each $h \in H$, and scan them one by one, terminating when the likelihood of one hypothesis is dominantly high.

Although a naive implementation of Algorithm 8 yields a running time that is linear in the number of actions, however since Score$(a; S)$ (Line 6 of Algorithm 8) can be calculated independently for each action a, one could paralyze this calculation for different actions and thus reducing the dependency on $|A|$. The same observation also holds for the rest algorithms to be introduced in the paper.

4.4.3 Proof Sketch for Proposition 9

We sketch a proof here and defer the details to § 4.9. The error analysis follows from standard concentration bounds, so we focus on the cost analysis. Suppose $\alpha > 0$, $\delta \in (0, 1/4]$, and $B \in (0, (1/2)\log\delta^{-1}]$. Let (σ^*, T^*) be any optimal partially adaptive algorithm, and let (σ, T) be the policy returned by RnB. Our analysis consists of the following steps:

(A) The sequence σ does well in covering the submodular functions, in terms of the total cover time: $\sum_{h\in H} \pi(h) \cdot \mathrm{CT}(f_h^B, \sigma) \le O\left(\log\left(|H|Bs^{-1}\right)\right) \sum_{h\in H} \pi(h) \cdot \mathrm{CT}(f_h^B, \sigma^*)$.

(B) The expected stopping time of our algorithm is not too much higher than the cover time of its submodular function: $\mathbb{E}_h[T] \le \alpha \cdot \mathrm{CT}(f_h^B, \sigma),\ \forall h \in H$.

(C) The expected stopping time in (σ^*, T^*) can be lower bounded in terms of the total cover time: $\mathbb{E}_h[T^*] \ge \Omega(s) \cdot \mathrm{CT}(f_h^B, \sigma^*),\ \forall h \in H$.

Proposition 9 follows by combining the above three steps. In fact,

$$\begin{aligned}\sum_{h\in H} \pi(h) \cdot \mathbb{E}_h[T] &\le \alpha \sum_{h\in H} \pi(h) \cdot \mathrm{CT}(f_h^B, \sigma)\\ &\le O\left(\alpha \log \frac{|H|B}{s}\right) \sum_{h\in H} \pi(h) \cdot \mathrm{CT}(f_h^B, \sigma^*)\\ &\le O\left(\frac{\alpha}{s} \log \frac{|H|B}{s}\right) \sum_{h\in H} \pi(h) \cdot \mathbb{E}_h[T^*],\end{aligned}$$

where $\sum_h \pi(h) \cdot \mathbb{E}_h[T]$ is the expected cost of our algorithm, and $\sum_h \pi(h) \cdot \mathbb{E}_h[T^*]$ is the expected cost of the optimal partially adaptive algorithm, $\mathrm{OPT}_\delta^{\mathrm{PA}}$.

At a high level, Step A can be showed by applying Theorem 14 and observing that the marginal positive increment of each f_h^B is $\Omega(s/(|H|B))$. Step B is implied by the correctness of the algorithm. In our key step, Step C, we fix an arbitrary δ-PAC-error partially adaptive algorithm (σ, T) and a hypothesis $h \in H$. Denote CT_h the cover time of f_h^B under permutation σ, with B chosen to be $\frac{1}{2}\log\delta^{-1}$, i.e., $\mathrm{CT}_h = CT(f_h^B, \sigma)$. Our goal is to lower bound $\mathbb{E}_h[T]$ in terms of CT_h. Given any

Algorithm 8 Partially Adaptive Algorithm: RnB(B, α)

```
Parameters: Coverage saturation level $B > 0$ and boosting intensity $\alpha > 0$.
Input: ASHT instance $(H, A, \pi, \mu, \mathcal{D})$
Initialize: $\sigma \longleftarrow \emptyset, \tilde{\sigma} \longleftarrow \emptyset$          ▷ Store the selected of actions.
for $t = 1, 2, ..., |\widetilde{A}|$ do          ▷ Rank: Compute a sequence of actions.
    $S \longleftarrow \{\sigma(1), ..., \sigma(t-1)\}$.          ▷ Actions selected so far.
    for $a \in \widetilde{A}$ do          ▷ Compute scores for each action.
        $$\mathrm{Score}(a; S) \longleftarrow \sum_{h \,:\, f_h^B(S) < 1} \pi(h) \frac{f_h^B(S \cup \{a\}) - f_h^B(S)}{1 - f_h^B(S)}.$$
    end for
      $\sigma(t) \longleftarrow \arg\max\{\mathrm{Score}(a; S) \;:\; a \in \widetilde{A} \backslash S\}$.          ▷ Select the greediest action.
end for
for  $t = 1, 2, ..., |\tilde{A}|$: do          ▷ Boost: Repeat each action in $\sigma$ for $\alpha$ times.
    for $i = 1, 2, ..., \alpha$: do
        $\tilde{\sigma}\big(\alpha(t-1) + i\big) \longleftarrow \sigma(t)$.
    end for
end for
for $t = 1, ..., \alpha|\tilde{A}|$: do
    Select action $\tilde{\sigma}(t)$ and observe outcome $y_t$.
    if  $t = \alpha \cdot \mathrm{CT}(f_h^B, \sigma)$ for some $h \in H$:  then          ▷ If $t$ is the timestamp for some $h$.
        for $g \in H \backslash \{h\}$: do
            $\Lambda(h, g) \longleftarrow \prod_{i=1}^{t} \mathbb{P}_{h,\tilde{\sigma}(i)}(y_i) / \mathbb{P}_{g,\tilde{\sigma}(i)}(y_i)$.          ▷ Compute the likelihood ratio.
        end for
        if $\log \Lambda(h, g) \geq \alpha B/2$ for all $g \in H \backslash \{h\}$, then
            return $h$.          ▷ Hypothesis identified.
        end if
    end if
end for
```

$d_1, ..., d_n$, denote $d^i = \sum_{j=1}^{i} d_j$. To this aim, we first show that for suitable choices of d_i's and t, the solution $z_i = \mathbb{P}_h[T = i]$ is feasible to the following LP:

$$\begin{aligned} LP(d,t): \quad & \min_z \sum_{i=1}^{N} i \cdot z_i \\ & s.t. \sum_{i=1}^{N} d^i z_i \geq \sum_{i=1}^{\mathrm{CT}_h - 1} d_i, \\ & \sum_{i=1}^{N} z_i = 1, \\ & z \geq 0. \end{aligned}$$

A feasible solution z can be viewed as a distribution of the stopping time. When $d_i = d(g, h; a_i)$, the first constraint says that the total KL-divergence "collected" at the stopping time has to reach a certain threshold. We show that $z_i = \mathbb{P}_h[T = i]$ is feasible, and the objective value of z is exactly $\mathbb{E}_h[T]$, hence $\mathbb{E}_h[T]$ is upper bounded by the LP-optimum $LP^*(d,t)$. Finally, we lower bound $LP^*(d, \mathrm{CT}_h - 1)$ by $\Omega(s \cdot \mathrm{CT}_h)$. The complete proof of Step C could be found in § 4.9.1.

4.5 Fully Adaptive Algorithm

In this section, we introduce our greedy fully adaptive algorithm. For ease of presentation, we only consider the scenario where the prior π is uniform over all hypotheses in this work. However, note that our guarantees hold for general priors. Our analysis is based on a reduction to the classical ODT problem.

4.5.1 Background: Optimal Decision Trees

In the ODT problem, an *unknown* true hypothesis h^* is drawn from a set of hypotheses H with some known probability distribution π. There is a set of known *tests*, each being a (deterministic) mapping from H to a finite *outcome space* set O. Thus, when performing a test, we can *rule out* the hypotheses that are inconsistent with the observed outcome, hence reducing the number of *alive* hypotheses. Moreover, the cost $c(T)$ of each test T is known, and the *cost of a decision tree* is defined to be the expected total cost of the tests selected until one hypothesis remains *alive*, in which case we say the true hypothesis is *identified*. The goal is to find a valid decision tree with minimal expected cost.

Note that the ODT problem can be viewed as a special case of the fully adaptive version of our problem where there is no noise and δ is 0. Consider the following greedy algorithm: let A be the alive hypotheses. Define Score(T) for each test T to be the minimal (over all possible outcomes)

Algorithm 9 **Fully Adaptive Algorithm**

```
1: Input: ASHT instance $(H, A, \pi, \mu, \mathcal{D})$ and error $\delta \in (0, 1/2)$.
2: $H_{\text{alive}} \longleftarrow H$.                                   ▷ Alive hypotheses.
3: while $|H_{\text{alive}}| \geq 2$ do
4:     $\hat{a} \longleftarrow \arg\max_{a \in A} \left\{ \min_{\omega \in \Omega_a} |H_{\text{alive}} \backslash T_a^\omega| \right\}$.    ▷ Greedy step.
5:     $c(\hat{a}) \longleftarrow \lceil s(\hat{a})^{-1} \log(|H|/\delta) \rceil$.    ▷ Num. of times to boost for sufficient confidence.
6:     Select $\hat{a}$ for $c(\hat{a})$ times consecutively and observe outcomes $X_1, ..., X_{c(\hat{a})}$.
7:     $\hat{\mu} \longleftarrow \sum_{i=1}^{c(\hat{a})} X_i$.                   ▷ Mean outcome.
8:     $\hat{\omega} \longleftarrow \arg\min\{|\hat{\mu} - \omega| : \omega \in \Omega_a\}$.    ▷ Round $\hat{\mu}$ to the closest $\omega$.
9:     $H_{\text{alive}} \longleftarrow H_{\text{alive}} \cap T_{\hat{a}}^{\hat{\omega}}$.    ▷ Update the alive hypotheses.
10: end while
```

number of alive hypotheses that it rules out in A. Then, we select the test T with the highest "bang-per-buck" Score(T)$/c(T)$. This algorithm is known to be an $O(\log|H|)$-approximation:

Theorem 15 (Chakaravarthy et al. 2009)**.** *For any ODT instance with uniform prior, the above greedy algorithm returns a decision tree whose cost is $O(\log|H|)$ times the optimum.*

4.5.2 Our Algorithm

We will analyze our greedy algorithm by relating to the above result. Consider the following ODT instance $\mathcal{I}_{\text{ODT}}$ for any given ASHT instance $\mathcal{I}$. The hypotheses set and prior in $\mathcal{I}_{\text{ODT}}$ are the same as in $\mathcal{I}$. For each action $a \in A$, let $\Omega_a := \{\mu(h, a) | h \in H\}$ be the mean outcomes. By Chernoff bound, we can show that when h is the true hypothesis, with high probability the mean outcome is "close" to $\mu(h, a)$ when a is repeated for $c(a)$ times. This motivates us to define a test $T_a : H \longrightarrow \Omega_a$ s.t. $T_a(h) = \mu(h, a)$, with cost $c(a) = \lceil s(a)^{-1} \log(|H|/\delta) \rceil$, where $s(a) = \min\{d(g, h; a) > 0 : g, h \in H\}$ is the separation parameter under action a. Such a test corresponds to selecting action a for $c(a)$ times consecutively in a row.

For each $\omega \in \Omega_a$, abusing the notation a bit, let $T_a^\omega \subseteq H$ denote the set of hypotheses whose outcome is ω when performing T_a, i.e., $T_a^\omega = \{h : \mu(h, a) = \omega\}$. At each step, Algorithm 9 selects an action $\hat{a}$ using the greedy rule (Step 4) and then repeat $\hat{a}$ for $c(\hat{a})$ times. Then we round the empirical mean of the observations to the closest element $\hat{\omega}$ in Ω_a, and rule out the hypotheses that are inconsistent with the observed outcome, i.e., the h's with $\mu(h, a) \neq \hat{\omega}$. We terminate when only one hypothesis remains alive.

4.5.3 Analysis

We sketch a proof for Theorem 12 and defer the details to § 4.10. Let h^* be the true hypothesis. By Hoeffding's inequality, in each iteration, with probability $1 - e^{-\log(|H|/\delta)} = 1 - \delta/|H|$ it holds $\hat{\omega} = \mu(h^*, \hat{a})$. Since in each iteration, $|H|$ decreases by at least 1, there are at most $|H| - 1$ iterations. Thus by union bound, the total error is at most δ.

Next we analyze the cost. Let GRE be the cost of Algorithm 9 and ODT* be the optimum of the ODT instance $\mathcal{I}_{\mathrm{ODT}}$. For the sake of analysis, we consider a "fake" cost $c' := \lceil s^{-1} \log(|H|/\delta) \rceil$, which does not depend on a. The definition of the ODT instance I_{ODT} remains the same except that each test has **uniform** cost c' (as opposed to $c(a)$). Let $c(T)$ and $c'(T)$ be the costs of the greedy tree T returned by Algorithm 2 under c and c' respectively. Then by Theorem 15, $c'(T) \le O(\log|H|) \cdot \mathrm{ODT}^*$. Note that $c' \le c(a)$ for each a since the separation parameter s is no larger than $s(a)$ by definition. Hence,

$$\mathrm{GRE} \le \mathrm{GRE} = c(T) \le c'(T) \le O(\log|H|) \cdot \mathrm{ODT}^*. \tag{4.1}$$

We relate ODT* to OPT_δ^{FA} using the following result (see proof in § 4.10):

Proposition 10. $\mathrm{ODT}^* \le O(s^{-1} \log(|H|/\delta)) \cdot \mathrm{OPT}_\delta^{FA}$.

The above is established by showing how to convert a δ-PAC-error fully adaptive algorithm to a valid decision tree, using only tests in $\{T_a\}$, and inflating the cost by a factor of $O(s^{-1} \log(|H|/\delta))$. Combining Proposition 10 with Eq. (4.1), we obtain

$$GRE \le O(s^{-1} \log \frac{|H|}{\delta} \log|H|) \cdot \mathrm{OPT}_\delta^{FA}.$$

Finally we remark that this analysis can easily be extended to general priors by reduction to the *adaptive submodular ranking* (ASR) problem (Navidi et al. 2020), which captures ODT as a special case. One may easily verify that the main theorem in Navidi et al. (2020) implies that a (slightly different) greedy algorithm achieves $O(\log(|H|))$-approximation for the ODT problem with general prior, test costs, and an arbitrary number of branches in each test. Thus for general prior, the same analysis goes through if we first reduce ASHT to ASR, and then replace the greedy step (Step 4 in Algorithm 9) with the greedy criterion for ASR.

4.6 Experiments

We performed a large set of numerical experiments, on both synthetic and real-world data (extending the analysis on the cancer genomic data from Cohen et al. (2018) described in the Introduction). Our results demonstrate the following:

1. Our algorithms can be applied to the liquid biopsy application, potentially yielding a cost that is substantially lower than the existing commercial panels and those constructed by state-of-art benchmarks.

2. Unlike existing benchmarks, our algorithms can explicitly account for prior information, yielding superior performance under more realistic priors.

3. Although our theoretic guarantees depend on the separability parameters s (which was introduced by the boosting steps), with small modifications our algorithms perform well when s is small on both synthetic and real-world data, outperforming state-of-art benchmarks.

Our benchmarks include two algorithms (one partially adaptive and one fully adaptive) based on a polynomial-time policy proposed by Naghshvar and Javidi (2013) (*Policy 1*[10]) and a completely random policy. The rest of this section is organized as follows: we first describe the benchmark policies and the implementation of our own policies. Then in § 4.6.2, we describe the setup and results of our synthetic experiments. Finally, in § 4.6.3, we test the performance of our algorithms on a publicly-available dataset of genetic mutations for cancer—COSMIC (Tate et al. 2019).

4.6.1 Algorithm Details

In all algorithms, we start with a uniform prior, and update our prior distribution (over the hypotheses space) each time an observation is revealed. Unless otherwise mentioned, the algorithm terminates if the posterior probability of a hypothesis is above the threshold $1 - \delta$. We first describe the random baseline and NJ's algorithms, and then discuss the modifications that we made to our algorithms.

Random Baseline At each step, an action was uniformly chosen from the set of all actions.

NJ's Algorithms *NJ Adaptive* Naghshvar and Javidi (2013) is a two-phase algorithm that solves a relaxed version of our problem, where the objective is to minimize a weighted sum of the expected number of tests and the likelihood of identifying the wrong hypothesis, i.e., $\min \mathbb{E}(T) + Le$, where T is the termination time, L is the penalty for a wrong declaration, and e is the probability of making that wrong declaration. The problem was formulated as a Markov decision process whose state space is the posterior distribution over the hypotheses. In Phase 1, which lasts as long as the posterior probability of all hypotheses is below a carefully chosen threshold, the action is sampled according to a distribution that is selected to maximize the minimum expected

[10] *Policy 2* in Naghshvar and Javidi (2013) does not have asymptotic guarantees and so is not considered in our experiments.

KL divergence among all pairs of outcome variables. In Phase 2, when one of the hypotheses has posterior probability above the chosen threshold, r, the action is sampled according to a distribution selected to maximize the minimum expected KL divergence between the outcome of this hypothesis and the outcomes of all other hypotheses. This threshold was optimized over in both synthetic and real-world experiments. The algorithm stops if the posterior of a hypothesis is above the threshold $1 - L^{-1}$. *NJ Partially Adaptive* contains only the Phase 1 policy.

Partially Adaptive In our synthetic experiments, we implement Algorithm 8 described in § 4.4 exactly, and set the boosting factor, α, to be 1. In our real-world experiments, we consider a variant of our algorithm. In particular, we consider the following modifications: 1) the amount of boosting is now a built-in feature of the algorithm, and 2) breaking ties according to some heuristic. Algorithm 10 formally describes our modified algorithm. To consider the amount of boosting as a built-in feature of the algorithm, we first generate a sequence of actions of length η for some large η value (with replacement) and then truncate the sequence to the minimum length to include all unique actions that have appeared in the sequence. When all actions in sequence σ has performed and we did not reach the target accuracy, then we repeat the entire sequence again. Our partially adaptive algorithm on COSMIC was generated by initializing η to be 800. Across all accuracy levels, the maximum truncated sequence length is 97.

Fully Adaptive We implement our algorithm described in § 4.5, with the modifications that 1) the amount of boosting is considered as a tunable parameter, 2) a hypothesis is only considered to be ruled out when we are deciding which action to perform, 3) we do not boost if no action can further distinguish any hypotheses in the alive set, 4) we break ties according to some heuristic. In particular, Modification 1) addresses the issues that our fully adaptive algorithm in § 4.5 overboosts. Modification b) controls the error probability δ when we decrease the amount of boosting. Modification c) handles small s without increasing the boosting factor. We formally describe this modified algorithm below.

Similar to NJ's algorithm, we maintain a probability distribution, ρ, over the set of hypotheses to indicate the likelihood of each hypothesis being the true hypothesis h^*. A hypothesis is considered to be ruled out at each step if the probability of that hypothesis is below a threshold in ρ. Throughout our experiments, we set this threshold to be $\delta/|H|$. At each step, after an action is chosen with certain repetitions and observation(s) is (are) revealed, we update ρ according to the realizations that we observed. Thus, under this setup, a hypothesis that was considered to be ruled out in the previous steps (due to "bad luck") could potentially become alive again.

At each iteration, for each action $a \in A$ and $k \in \mathbf{N}$, we define $T_{a,k}$ to be the meta-test that repeats action a for k times consecutively, and we define its cost to be $c(T_{a,k}) = kc_a$. By Chernoff bound,

Algorithm 10 Partially Adaptive Algorithm in the COSMIC Experiment

```
Parameters: Coverage saturation level $B > 0$ and maximum sequence length $\eta > 0$.
Input: ASHT instance $(H, A, \pi, \mu)$
Output: actions sequence $\sigma$
Initialize: $\sigma \longleftarrow \emptyset$    ▹ Store the selected of actions
for $t = 1, 2, ..., \eta$ do do    ▹ Rank: Compute a sequence of actions of length $\eta$
    $S \longleftarrow \{\sigma(1), ..., \sigma(t-1)\}$.    ▹ Actions selected so far
    for  $a \in A$ do    ▹ Compute scores for each action
        $$\text{Score}(a; S) \longleftarrow \sum_{h : f_h^B(S) < 1} \pi(h) \frac{f_h^B(S \cup \{a\}) - f_h^B(S)}{1 - f_h^B(S)}.$$
    end for
    $\sigma(t) \longleftarrow \arg\max\{\text{Score}(a; S) : a \in A\}$.    ▹ Select the greediest action and break ties
    according the heuristic described in Algorithm 11
end for
Let i be the largest index for which the an action appears the first time in sequence $\sigma$, then
we return the sequence $(\sigma(1), ...., \sigma(i))$.
```

with k i.i.d. samples, we may construct a confidence interval of width $\sim k^{-1/2}$. This motivates us to rule out the following hypotheses when $T_{a,k}$ is performed. Let $\bar{\mu}$ be the observed mean outcome of these k samples, we define the elimination set to be $E_{\bar{\mu}}(T_{a,k}) := \{h : |\mu(h, a) - \bar{\mu}| \geq Ck^{-1/2}\}$, where C is set to be 1/2 in our uniform prior experiment and set to be 1/3 in our non-uniform prior experiment. To define greedy, we need to formalize the notion of bang-per-buck. Suppose H_{alive} is the current set of alive hypotheses. We define the score of a test as the number of alive hypotheses ruled out in the worst-case over all possible mean outcomes $\bar{\mu}$. Formally, the score of $T_{a,k}$ w.r.t mean outcome $\bar{\mu}$ is

$$\text{Score}_{\bar{\mu}}(T_{a,k}) = \text{Score}_{\bar{\mu}}(T_{a,k}; H_{alive}) = \frac{|E(T_{a,k}; \bar{\mu}) \cap H_{alive}|}{c(T_{a,k})},$$

and define its worst-case score to be $\text{Score}(T_{a,k}) = \min\{\text{Score}_{\bar{\mu}}(T_{a,k}) : \bar{\mu} \in \{0, 1/k, ..., 1\}\}$.

Our greedy policy simply selects the test T with the highest score, formally, select $T_{a,k} = \arg\max\{\text{Score}(T) : k \leq k_{\max}, a \in A\}$.

In the synthetic experiments, we set $k_{\max} = 5$. In the real-world experiments, we consider the cases where $k \in \{15, 20, 25, 30\}$ (with $k_{\max} = 30$) when the prior is uniform and $k \in \{5, 10, 15, 20\}$ (with $k_{\max} = 20$) when the prior is non-uniform. If several actions have the same greedy score, then we choose the action a^* whose sum of the KL divergence of pairs of $\mu(h, a^*)$ is the largest, and breaking ties arbitrarily. If no action can further distinguish any hypotheses in the alive set, then we set the boosting factor to be 1 and use the above heuristic to choose the action to perform. The algorithm is formally stated in Algorithm 11.

Algorithm 11 Adaptive experiments: FA($k_{\max}, \delta$)

```
Parameters: maximum boosting factor $k_{\max} > 0$ and convergence threshold $\delta > 0$
Input: ASHT instance $(H, A, \pi, \mu)$, current posterior of the true hypothesis vector $\rho$
Output: the test $T_{a,k}$ to perform in the next iteration
Let $H_{\text{alive}}$ be the set of hypotheses $i$ whose posterior probability $\rho_i$ is above $\delta/|H|$.
for $k = 1, 2, ..., k_{\max}$ do
    For each $a \in \widetilde{A}$ define:
        $$\text{Score}_{\bar{\mu}}(T_{a,k}) = \text{Score}_{\bar{\mu}}(T_{a,k}; H_{\text{alive}}) = \frac{|E(T_{a,k}; \bar{\mu}) \cap H_{\text{alive}}|}{c(T_{a,k})},$$
    where $E_{\bar{\mu}}(T_{a,k}) := \{h : |\mu(h, a) - \bar{\mu}| \geq Ck^{-1/2}\}$, and $c(T_{a,k}) = kc_a$. We define the worst-case
    score of a test to be:
        $$\text{Score}(T_{a,k}) = \min\{\text{Score}_{\bar{\mu}}(T_{a,k}) : \bar{\mu} \in \{0, 1/k, ..., 1\}\}.$$
end for
Compute greediest action
    $$G = \arg\max\{\text{Score}(T) : k \leq k_{max}, a \in A\}.$$
if the Score of each test in $G$ equals to 0, i.e, no test can further distinguish between the alive
    hypotheses under $k_{\max}$ then
    we choose the action $a^*$ such that $a^* = \arg\max \sum_{h,g \in H_{\text{alive}}} \text{KL}(\mu(h, a), \mu(g, a))$, breaking ties
    randomly, and return $k = 1$.
else
    if $G$ is a singleton, then we return $G$. Else, we choose the action $a^*$ such that $a^* =$
    $\arg\max_G \sum_{h,g \in H_{\text{alive}}} \text{KL}(\mu(h, a), \mu(g, a))$, and breaking ties randomly.
end if
```

4.6.2 Synthetic Experiments

Parameter Generation and Setup Fig. 4.2 summarizes the results of our partially and fully adaptive experiments on synthetic data. Both figures were generated with 100 instances: each with 25 hypotheses and 40 actions. The outcome of each action under each hypothesis is binary, i.e., the $D_{\mu(h,a)}$'s are the Bernoulli distributions, where $\mu(a, h)$ were *uniformly* sampled from the

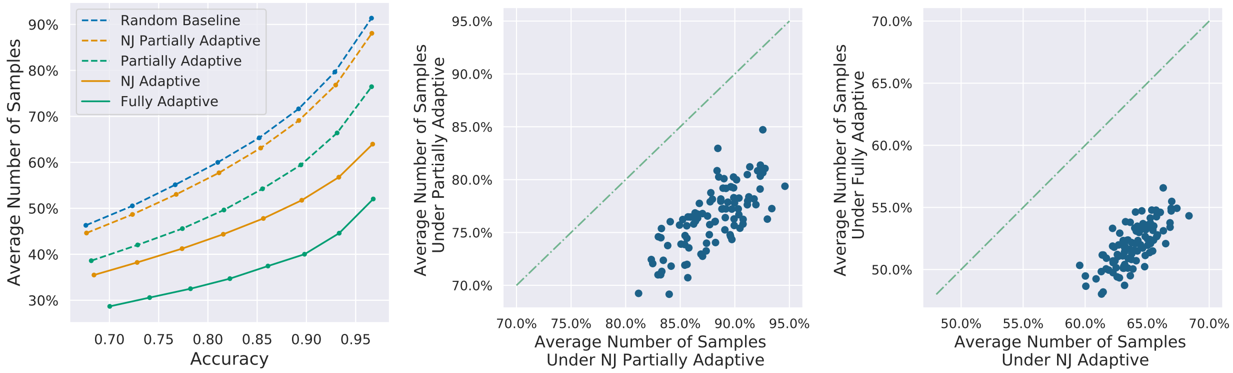


Figure 4.2: Comparison of our fully and partially adaptive algorithms with *NJ Adaptive*, *NJ Partially Adaptive* and *Random Baseline* on synthetic data. The average number of samples is normalized with respect to the largest number of sample required in *Random Baseline*. Left: each dot corresponds to the average performance of 100 randomly generated instances each averaged over 2,000 replications. Middle and Right: contains the same 100 instances in the left figure. Each dot corresponds to one instance and each averaged over 2,000 replications. Middle and Right: the average accuracies of those 100 instances in all algorithms equal to 0.97.

[0,1] interval. Each instance was then averaged over 2,000 replications, where a "ground truth" hypothesis was randomly drawn. The prior distribution, π, was initialized to be uniform for all runs. On the horizontal axis, the accuracies of both algorithms were averaged over these 100 instances, where the accuracy is calculated as the percentage of correctly identified hypotheses among the 2,000 replications. On the vertical axis, the number of samples used by the algorithm is first averaged over the 2,000 replications and then averaged over the 100 instances.

Results In Fig. 4.2 (left), we observe that 1) the performance of our fully adaptive algorithm dominates those of all other algorithms, 2) our partially adaptive algorithm outperforms all other partially adaptive algorithms, and 3) the performance of adaptive algorithms outperform those of partially adaptive algorithms. The threshold for entering Phase 2 policy in *NJ Adaptive* was set to be 0.1. Indeed, we observe that *NJ Adaptive* outperforms *NJ Partially Adaptive*. In Fig. 4.2 (middle), δ equals to 0.05 for both *NJ Partially Adaptive* and *Partially Adaptive*. We observe that our partially and fully adaptive algorithms outperform *NJ Partially Adaptive* and *NJ Adaptive* instance-wise by large margins respectively in Fig. 4.2 middle and left.

4.6.3 Real-World Experiments

Problem Setup Our real-world experiment is motivated by the design of DNA-based blood tests to detect cancer. In such a test, genetic mutations serve as potential signals for various cancer types, but DNA sequencing is, even today, expensive enough that the 'amount' of DNA that can be sequenced in a single test is limited if the test is to remain cost-effective. For example, one of the most-recent versions of these tests Cohen et al. (2018) involved sequencing just 4,500 *addresses* (from among 3 billion total addresses in the human genome), and other tests have had similar scale (e.g., Razavi et al. 2017, Chan et al. 2017, Phallen et al. 2017). Thus, one promising approach to the ultimate goal of a cost-effective test is adaptivity.

Our experiments are a close reproduction of the setup used by Cohen et al. (2018) to identify their 4,500 addresses. We use genetic mutation data from real cancer patients—the publicly-available COSMIC (Tate et al. 2019, Cosmic 2019) dataset, which includes the de-identified gene-screening panels for 1,350,015 patients. We treated 8 different types of cancer (as indicated in Cohen et al. 2018) as the 8 hypotheses, and identified 1,875,408 potentially mutated genetic addresses. To extract the tests, we grouped the the genetic addresses within an interval of 45 (see Cohen et al. 2018 for the biochemical reasons behind this choice), resulting in 581,754 potential tests. We then removed duplicated tests (i.e., the tests that share the same outcome distribution for all 8 cancer types), resulting in 23,135 final tests that we consider in our experiments. From the data, we extracted a "ground-truth" table of mutation probabilities containing the likelihood of a mutation in any of the 23,135 genetic address intervals being found in patients with any of the 8 cancer types. This served as the instance for our experiment. The majority of the mutation probabilities in our instances was either zero or some small positive number. To calculate the KL divergence between these probabilities, we replace zero with the number 10^{-10} in our instance.

Uniform Prior Results Although in reality, all patients have different priors for having different cancers, in our first set of experiments, to demonstrate the advantage of our algorithms, we assume that the truth hypothesis (cancer type) was drawn uniformly, and we initialize uniform priors for all algorithms. Fig. 3.6, summarizes the performance of our algorithms under the assumption of uniform prior. Similar to Fig. 4.2, in Fig. 3.6 (top row left) we observe first that the performance of our algorithm dominates those to the rest algorithms. We also observe that our partially adaptive algorithm outperforms *NJ Partially Adaptive*. However, unlike Fig. 4.2, we observe that *NJ Adaptive* underperforms *Partially Adaptive* when the accuracies are low on this instance. The threshold for entering Phase 2 policy, r, in *NJ Adaptive* was set to be 0.3. Since Phase 1 policy is less efficient than Phase 2 policy, we observe that the performance of *NJ Adaptive* is convex with respect to r—when r is small, the algorithm is more likely to alternate between Phases 1 and 2 policies and when r is large we spend more time in Phase 1 policy. As a result, we observe that variance of *NJ Adaptive* is relatively high when compared with those of our algorithms. On the other hand, we observe that our fully adaptive algorithm enjoys a narrower confidence interval as well as a better performance. Note that due to the nature of the sparsity of our instance, the performance of the random baseline was very poor when compared with these of *NJ Adaptive* and *Fully Adaptive* and thus was excluded. Fig. 3.6 (top middle) is the confusion matrix corresponding to our fully adaptive algorithm where the algorithm accuracy equals to 0.97, and Fig. 3.6 (top right) corresponds to the sensitivity and specificity of our algorithm for each cancer type (in the same ordering as) in the top middle figure. We observe that our adaptive algorithm performs reasonably well across different underlying true hypotheses (i.e., different cancer types). The middle and bottom rows of Fig. 3.6 contain the sensitivity figures for each cancer type that are analogous to Fig. 4.1. Due to the number of samples for each cancer type is limited, these plots are more volatile than Fig. 3.6 top left. We observe that our fully adaptive algorithm outperforms the rest algorithms for the majority of cancer types. In addition, for those cancer types that our fully algorithm underperforms (e.g., breast cancer), we show below that when the prior of having a particular cancer increases, our algorithm becomes better at identifying this cancer.

Non-Uniform Prior Results To mimic the distribution of different cancer types in the real-world population, we exacted the number of newly diagnosed cancer cases from cancer.org (2021) to form the priors in our algorithms (the prior information is included in Fig. 4.4 bottom right). Note that neither *NJ Partially Adaptive* nor *NJ Adaptive* takes the prior into account when constructing the action sequences (or policies when randomization is allowed). Thus, the Phases 1 and 2 policies remain the same. (The prior was taken into account in the exiting criterion of both algorithms, and in addition, was taken into account in *NJ Adaptive* when entering Phase 2 policies.) On the other hand, both our partially adaptive and fully adaptive algorithms take the prior into the account when constructing the action sequences, i.e., the action sequences produced by our algorithms change as the prior changes. Thus, we expect our algorithms to outperform their partially or fully adaptive counterparts under the non-uniform prior as well. Indeed, Fig. 4.4 summarizes the results, and we observe that our partially adaptive and fully adaptive algorithms dominate NJ's partially and fully adaptive algorithms by a large margin, respectively. The threshold for entering Phase 2 policy, r, in *NJ Adaptive* was increased to 0.45 under this prior, and similar to Fig. 3.6, we observe that the variance of *NJ Adaptive* is high when compared with our algorithms, indicating the trade off between variance and performance. Similar to Fig. 3.6, Fig. 4.4 (top middle) is the confusion matrix corresponding to our fully adaptive algorithm where the realized algorithm accuracy equals to 0.97. Fig. 4.4 (top right) contains the sensitivity, specificity, and the prior of our algorithm for each cancer type in the top middle figure. Similarly, we include the sensitivity figures for each cancer type in Fig. 4.4 middle and bottom rows. When compared with Fig. 3.6, we observe that when the weight of a cancer type increases, with tuning, our algorithm becomes better at identifying those cancers (e.g., breast cancer and lung cancer) while maintaining the accuracies in identifying the rest cancer types.

4.7 Conclusion

In this work, we studied problem of active sequential hypothesis testing, motivated particularly by the design of adaptive liquid biopsies. We provided the first approximation guarantees for the ASHT problem in both the partially adaptive and fully adaptive setting, which grows linearly in the separability parameter s^{-1} and logarithmically in the number of candidate hypotheses and the (inverse) error rate δ^{-1}. Moreover for the partially adaptive version, by combining the SFR framework with a novel LP-based analysis, we improved the dependence on δ from $\log \frac{1}{\delta}$ in the naive analysis to $\log\log \frac{1}{\delta}$, which is much more favorable since in practice δ is usually very small. We further extend the fully adaptive algorithm to the total-error version by introducing a novel chance-constrained ODT problem (§ 4.11).

To illustrate the applicability of our proposed method to the liquid biopsy problem, we conducted numerical studies on the COSMIC dataset. We found that our algorithms outperform the existing state-of-art benchmarks by large margins. Furthermore, our algorithms consider the priors for having different cancer types explicitly when constructing the action sequences, yielding superior performances under non-uniform priors. Finally, although the theoretical guarantees of our algorithms depend on the separability parameter s, we showed numerically that our modified algorithms work well in practice on both synthetic and real-world data.

4.8 Assumption: Log Likelihood Ratio Between Two Hypotheses

To show the correctness of our algorithm, we need to consider the *log-likelihood ratio* (LLR), formally defined as follows:

Definition 7. *For any $a \in A$ and $h, g \in H$, define $Z(h, g; a) = \log \frac{\mathbb{P}_{h,a}(\xi)}{\mathbb{P}_{g,a}(\xi)}$ where $\xi \sim D_{\mu(h,a)}$.*

We will assume that the subgaussian norm of the LLR between two hypotheses is not too large when compared to the difference of their parameters, as formalized below.

Definition 8. *Let $\rho > 0$ be the minimal number s.t. for any distinct pair of hypotheses $h, g \in H$ and action $a \in A$, it holds that $\|Z(h, g; a)\|_{\psi_2} \leq \rho \cdot |\mu(g, a) - \mu(h, a)|$.*

We will present an error analysis for general ρ. Prior to that, we first point out that many common distributions satisfy $\rho = O(1)$.

Examples It is straightforward to verify that $\rho = O(1)$ for the following common distributions:

- Bernoulli distributions: $D_\theta = Ber(\theta)$ where $\theta \in [\theta_{min}, \theta_{max}]$ for constants $\theta_{min}, \theta_{\max} \in (0, 1)$, and

- Gaussian distributions: $D_\theta = N(\theta, 1)$ where $\theta \in [\theta_{min}, \theta_{max}]$ for constants $\theta_{min} < \theta_{\max}$.

Take Bernoulli distribution as an example. Fix any hypotheses $h, g \in H$ and action $a \in A$, write $\Delta = \mu(h, a) - \mu(g, a)$. Then, $Z = Z(h, g; a)$ can be rewritten as

$$Z = \begin{cases} \log(1 + \frac{\Delta}{\mu(g,a)}), & \text{w.p. } \mu(h, a), \\ \log(1 - \frac{\Delta}{1-\mu(g,a)}), & \text{w.p. } 1 - \mu(h, a). \end{cases}$$

Since $0 < \theta_{min} \leq \mu(g, a) \leq \theta_{\max} < 1$, we have $|Z| \leq C|\Delta|$ almost surely where $C = 2\max\{(1 - \theta_{max})^{-1}, \theta_{min}^{-1}\}$. Moreover, it is known that (see Vershynin 2018) any subgaussian random variable Z satisfies $\|Z\|_{\psi_2} \leq \frac{1}{\ln 2}\|Z\|_\infty$, so it follows that

$$\|Z\|_{\psi_2} \leq \frac{1}{\ln 2}\|Z\|_\infty \leq \frac{C\Delta}{\ln 2} = O(\Delta).$$

Thus $\rho = O(1)$.

4.9 Proof of Proposition 9

4.9.1 Error Analysis

We first prove that at each timestamp $\tau(h)$, with high probability our algorithm terminates and returns h.

Lemma 9. *Let $B > 0$. If $h \in H$ is the true hypothesis, then w.p. $1 - e^{-\Omega(\rho^{-2}\alpha B)}$, it holds $\log \Lambda(h, g; \tau(h)) \geq \frac{1}{2}\alpha B$ for all $g \neq h$.*

Proof. Proof of Lemma 9 Let $\tilde{\sigma} = (a_1, a_2, ...)$ be the sequence *after* the boosting step, so $a_1 = ... = a_\alpha, a_{\alpha+1} = ... = a_{2\alpha}$, so on so forth. Write $Z_i = Z(h, g; a_i)$, then for any $t \geq 1$, it holds $\log \Lambda(h, g; t) = \sum_{i=1}^{t} Z_i$. By the definition of cover time, $\sum_{i=1}^{\tau(h)} d(h, g; a_i) = \sum_{i=1}^{\tau(h)} \mathbb{E}[Z_i] \geq \alpha B$. Thus,

$$\begin{aligned} \mathbb{P}_h \left[\log \Lambda\big(h, g; \tau(h)\big) < \frac{1}{2}\alpha B\right] &= \mathbb{P}_h \left[\sum_{i=1}^{\tau(h)} Z_i < \frac{1}{2}\alpha B\right] \\ &\leq \mathbb{P}_h \left[\left|\sum_{i=1}^{\tau(h)} Z_i - \sum_{i=1}^{\tau(h)} \mathbb{E}Z_i\right| > \frac{1}{2}\sum_{i=1}^{\tau(h)} \mathbb{E}Z_i\right]. \end{aligned} \tag{4.2}$$

By Theorem 1,

$$\text{Eq.}(4.2) \leq \exp\left(-\Omega\left(\frac{(\alpha B)^2}{\sum_{i=1}^{\tau(h)} \|Z_i\|_{\psi_2}^2}\right)\right). \tag{4.3}$$

We next show that $\sum_{i=1}^{\tau(h)} \|Z_i\|_{\psi_2}^2 \leq O(\rho^2 \alpha B)$. Write $\Delta_i = \mu(h, a_i) - \mu(g, a_i)$, then by Assumption 3, $\Delta_i^2 \leq C_2 \cdot d(h, g; a_i)$. Note that $\|Z_i\|_{\psi_2} \leq \rho\Delta_i$, so it follows that

$$\sum_{i=1}^{\tau(h)} \|Z_i\|_{\psi_2}^2 \leq \rho^2 \sum_{i=1}^{\tau(h)} \Delta_i^2 \leq C_2\rho^2 \sum_{i=1}^{\tau(h)} d(h, g; a_i). \tag{4.4}$$

Recall that σ is the sequence *before* boosting. Write $t = CT(f_h^B, \sigma)$ for simplicity. By the definition of cover time,

$$\sum_{i=1}^{\alpha t} d(h, g; a_i) \geq \alpha B \geq \sum_{i=1}^{\alpha(t-1)} d(h, g; a_i).$$

Note that $\tau(h) = \alpha t$, so

$$\sum_{i=1}^{\alpha t} d(h, g; a_i) \leq 2 \sum_{i=1}^{\alpha(t-1)} d(h, g; a_i) \leq 2\alpha B.$$

Combining the above with Eq. (4.4), we have

$$\sum_i \|Z_i\|_{\psi_2}^2 \leq 2C_2\rho^2 \alpha B.$$

Substituting into Eq. (4.3), we obtain

$$\mathbb{P}_h\left[\log\Lambda\big(h,g;\tau(h)\big) < \frac{1}{2}\alpha B\right] \le e^{-\Omega(\rho^{-2}\alpha B)}.$$

The proof completes by applying the union bound over all $g \in H\backslash\{h\}$. ■ □

By a similar approach we may also show that it is unlikely that the algorithm terminates at a wrong time stamp before scanning the correct one.

Lemma 10. *Let $B > 0$. If $h \in H$ is the true hypothesis, then for any $g \neq h$, it holds that $\log\Lambda(g,h;\tau(g)) < \frac{1}{2}\alpha B$ with probability $1 - e^{-\Omega(\rho^{-2}\alpha B)}$.*

We are able to bound the error of the RnB algorithm by combining Lemma 9 and Lemma 10.

Proposition 11. *For any true hypothesis $h \in H$, algorithm RnB(B,α) returns h with probability at least $1 - |H|e^{-\Omega(\rho^{-2}\alpha B)}$. In particular, if the outcome distribution D_μ is Ber(μ), then $\rho = O(1)$ and the above probability becomes $1 - |H|e^{-\Omega(\alpha B)}$.*

4.9.2 Cost Analysis

Recall that in § 4.4, only Step (C) remains to be shown, which we formally state below.

Proposition 12. *Let (σ, T) be a δ-PAC-error partially adaptive algorithm. For any $B \le \log\delta^{-1}$ and $h \in H$, it holds that $\mathbb{E}_h[T] \ge \Omega\big(s \cdot \mathrm{CT}(f_h^B, \sigma)\big)$.*

We fix an arbitrary $h \in H$ and write $\mathrm{CT}_h := \mathrm{CT}(f_h^B, \sigma)$, where we recall that σ is the sequence of actions before boosting (do not confuse with $\tilde{\sigma}$). To relate the stopping time T (under h) to the cover time of the submodular function for h in σ, we introduce a linear program. We will show that for suitable choice of d, we have

- $LP^*(d, \mathrm{CT}_h - 1) \le \mathbb{E}_h[T]$, and
- $LP^*(d, \mathrm{CT}_h - 1) \ge \Omega(s \cdot \mathrm{CT}_h)$.

Hence proving Step (C) in the high-level proof sketched in § 4.4.

We now specify our choice of d. For any $d_1, ..., d_N \in \mathbb{R}_+$, write $d^t := \sum_{i=1}^t d_i$ for any t and consider

$$\begin{aligned} LP(d,t):\quad & \min_z \sum_{i=1}^N i \cdot z_i \\ & s.t.\ \sum_{i=1}^N d^i z_i \ge d^t, \\ & \sum_{i=1}^N z_i = 1, \\ & z \ge 0. \end{aligned}$$

We will consider the following choice of d_i's. Suppose (σ, T) has δ-PAC-error where $\delta \in (0, 1/4]$. For any pair of hypotheses h, g and any set of actions S, define

$$K^B_{h,g}(S) = \min\left\{1, B^{-1}\sum_{a\in S} d(h, g; a)\right\}.$$

Hence,

$$f^B_h(S) = \frac{1}{|H| - 1}\sum_{g\in H\backslash\{h\}} K^B_{h,g}(S).$$

Fix any $B \le \log \delta^{-1}$ and let g be the last hypothesis separated from h, i.e.,

$$g \ := \arg\max_{h'\in H\backslash\{h\}} \left\{\mathrm{CT}(K^B_{h,h'}, \sigma)\right\}.$$

Then by the definition of cover time, we have $\mathrm{CT}_h = \mathrm{CT}(f^B_h, \sigma) = \mathrm{CT}(K^B_{hg}, \sigma)$. Without loss of generality[11], we assume all actions a satisfy $\mu(h, a) = \mu(g, a)$ in $\tilde{\sigma} = (a_1, .., a_N)$. We choose the LP parameters to be $d_i = d(h, g, a_i)$ for $i \in [N]$.

Outline We will first show that the LP optimum is upper bounded by the expected termination time T (Proposition 13). We then lower bound it in terms of CT_h (Proposition 14).

Proposition 13. *Suppose (σ, T) has δ-PAC-error for some $0 < \delta \le \frac{1}{4}$. Let $z_i = \mathbb{P}_h[T = i]$ for $i \in [N]$, then $z = (z_1, ..., z_N)$ is feasible to LP$(d, \mathrm{CT}_h - 1)$.*

Note that $\mathbb{E}_h[T]$ is simply the objective value of z, thus Proposition 13 immediately implies:

Corollary 2. $\mathbb{E}_h[T] \ge LP^*(d, \mathrm{CT}_h - 1)$.

We next lower bound the expected log-likelihood when the algorithm stops.

Lemma 11 (Nowak 2009)**.** *Let $\mathbb{A}$ be any algorithm (not necessarily partially adaptive) for the ASHT problem. Let $h, g \in H$ be any pair of distinct hypotheses and O be the random output of $\mathbb{A}$. Define the error probabilities $P_{hh} = \mathbb{P}_h(O = h)$ and $P_{hg} = \mathbb{P}_h(O = g)$. Let Λ be the likelihood ratio between h and g when $\mathbb{A}$ terminates. Then,*

$$\mathbb{E}_h[\log\Lambda] \ge P_{hh}\log\frac{P_{hh}}{P_{hg}} + (1 - P_{hh})\log\frac{1 - P_{hh}}{1 - P_{hg}}.$$

Proof. Proof of Lemma 11 Let $\mathcal{E}$ be the event that the output is h. Then by Jensen's inequality,

$$\mathbb{E}_h[\log\Lambda_T|\mathcal{E}] \ge -\log\mathbb{E}_h[\Lambda^{-1}|\mathcal{E}] = -\log\frac{\mathbb{E}_h[\mathbb{1}(\mathcal{E})\cdot\Lambda^{-1}]}{\mathbb{P}_h(\mathcal{E})}. \tag{4.5}$$

Recall that an algorithm can be viewed as a decision tree in the following way. Each internal node is labeled with an action, and each edge below it corresponds to a possible outcome; each

[11] If there is some action a with $d(h, g; a) = 0$, then we simply remove it. This will not change the argument.

leaf corresponds to termination, and is labeled with a hypothesis corresponding to the output. Write $\sum_x$ as the summation over all leaves and let $p_h(x)$ (resp. $p_g(x)$) be the probability that the algorithm terminates in leaf x under h (resp. g), then,

$$\begin{aligned}\mathbb{E}_h[\mathbb{1}(\mathcal{E})\cdot\Lambda^{-1}] &= \sum_x \mathbb{1}(x\in\mathcal{E})\cdot\Lambda^{-1}(x)\cdot p_h(x)\\ &= \sum_x \mathbb{1}(x\in\mathcal{E})\cdot\frac{p_g(x)}{p_h(x)}p_h(x)\\ &= \sum_x \mathbb{1}(x\in\mathcal{E})\cdot p_h(x)\\ &= \mathbb{E}_h[\mathbb{1}(x\in\mathcal{E})] = P_{hg}.\end{aligned}$$

Combining the above with Equation (4.5), we obtain

$$\mathbb{E}_h[\log\Lambda|\mathcal{E}] \geq \log\frac{P_{hh}}{P_{hg}}.$$

Similarly, we have $\mathbb{E}_h(\log\Lambda|\bar{\mathcal{E}}) \geq \log\frac{1-P_{hh}}{1-P_{hg}}$, where $\bar{\mathcal{E}}$ is the event that the output is not h. The proof follows immediately by combining these two inequalities. ■ □

To show Proposition 13 we need a standard concept—*stopping time.*

Definition 9 (Stopping time Mitzenmacher and Upfal 2017)**.** *Let $\{X_i\}$ be a sequence of random variables and T be an integer-valued random variable. If for any integer t, the event $\{T = t\}$ is independent with $X_{t+1}, X_{t+2}, \ldots$, then T is called a* ***stopping time*** *for X_i's.*

Lemma 12 (Wald's Identity)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be independent random variables with means $\{\mu_i\}_{i\in\mathbb{N}}$, and let T be a stopping time w.r.t. X_i's. Then, $\mathbb{E}[\sum_{i=1}^T X_i] = \mathbb{E}[\sum_{i=1}^T \mu_i]$.*

Proof of Proposition 13. One may verify that the lower bound in Lemma 11 is increasing w.r.t P_{hh} and decreasing w.r.t P_{hg}. Therefore, since $\mathbb{A}$ has δ-PAC-error, by Lemma 11 it holds that

$$\mathbb{E}_h[\log\Lambda(h,g;T)] \geq (1-\delta)\log\frac{1-\delta}{\delta} + \delta\log\frac{\delta}{1-\delta} \geq \frac{1}{2}\log\frac{1}{\delta} \geq B \geq d^{\mathrm{CT}_h - 1}.$$

By Lemma 12,

$$\sum_{i=1}^N d^i z_i = \sum_{i=1}^N d^i\cdot\mathbb{P}_h(T=i) = \mathbb{E}_h[\log\Lambda(h,g;T)].$$

The proof follows by combining the above. ■

So far we have upper bounded $LP^*(d, \mathrm{CT}_h - 1)$ using $\mathbb{E}_h[T]$. To complete the proof, we next lower bound $LP^*(d, \mathrm{CT}_h - 1)$ by $\Omega(s\cdot\mathrm{CT}_h)$.

Lemma 13. *$LP^* = \min_{i\leq t<j} LP^*_{ij}$ where $LP^*_{ij} = i + (j-i)\frac{d^t - d^i}{d^j - d^i}$.*

Proof. **Proof of Lemma 13** Observe that for any optimal solution, the inequality constraint must be tight. By linear algebra, we deduce that any basic feasible solution has support size two.

Consider the solutions whose only nonzero entries are i, j. Then, $LP(d,t)$ becomes

$$\begin{aligned} LP_{ij}(d,t): \quad \min_{z_i,z_j} \quad & iz_i + jz_j \\ s.t.\ & d^i z_i + d^j z_j = d^t, \\ & z_i + z_j = 1, \\ & z \geq 0. \end{aligned}$$

Note that since $d^i < d^j$, $LP_{i,j}(d,t)$ admits exactly one feasible solution, whose objective value can be easily verified to be $LP^*_{ij} := i + (j-i)\frac{d^t - d^i}{d^j - d^i}$. ■ □

Now we are ready to lower bound the LP optimum.

Proposition 14. *For any $d = (d_1, ..., d_N) \in \mathbb{R}^N$ and $t \in \mathbb{N}$, it holds that $LP^*(d,t) \geq t \cdot \min\{d_i\}_{i\in[N].}$*

Proof. **Proof of Lemma 14** By Lemma 13, it suffices to show that $LP^*_{ij} \geq d^t$ for any $i \leq t < j$. Since $d^k < k$ for any integer k,

$$(j - d^t)(d^t - d^i) \geq (d^j - d^t)(d^t - i).$$

Rearranging, the above becomes

$$i(d^j - d^i) + (j-i)(d^t - d^i) \geq d^t(d^j - d^i),$$

i.e.,

$$i + (j-i)\frac{d^t - d^i}{d_j - d^i} \geq d^t.$$

Note that the LHS is exactly LP^*_{ij}, thus $LP^*(d,t) \geq d^t \geq t \cdot \min\{d_i\}_{i\in[N]}$ for any $t \in \mathbb{N}$. ■ □

It immediately follows that $LP^*(d,t) \geq st$, completing the proof of Proposition 9.

4.10 Proof of Proposition 10

We first formally define a decision tree, not only for mathematical rigor but more importantly, for the sake of introducing a novel variant of ODT. Recall that Ω is the space of the test outcomes, which we assume to be discrete for simplicity.

Definition 10 (Decision Trees). *A decision tree is a rooted tree, each of whose interior (i.e., non-leaf) node v is associated with a state (A_v, T_v), where T_v is a test and $A_v \subseteq H$. Each interior node has $|\Omega|$ children, each of whose edge to v is labeled with some outcome. Moreover, for any interior node v, the set of alive hypotheses A_v is the set of hypotheses consistent with the outcomes on the edges of the* *path from the root to v. A node ℓ is a leaf if $|A_\ell| = 1$. The decision tree terminates and outputs the only alive hypothesis when it reaches a leaf.*

To relate OPT_δ^{FA} to the optimum of a suitable ODT instance, we introduce a novel variant of ODT. As opposed to the ordinary ODT where the output needs to be correct with probability 1, in the following variant, we consider decision trees which may *err* sometimes:

Definition 11 (Incomplete Decision Trees)**.** *An incomplete decision tree is a decision tree whose leaves ℓ's are associated with* states *(A_ℓ, p_ℓ)'s, where A_ℓ represents the subset of hypotheses consistent with all outcomes so far, and p_ℓ is a distribution over A_ℓ. A hypothesis is randomly drawn from p_ℓ and is returned as the identified hypothesis (possibly wrong).*

Now we already to introduce *chance-constrained ODT problem* (CC-ODT). Given an error budget $\delta > 0$, we aim to find the minimal cost decision tree whose error is within δ. There are two natural ways to interpret "error", which will both be considered in § 4.10 and § 4.11. In the first one, we require the error probability under *any* hypothesis to be lower than the given error budget. In the other one, we only require the *expected* error probability over all hypotheses to be within the budget. Intuitively, the second version allows for more flexibility since the errors under different hypotheses may differ significantly, rendering the analysis more challenging since we do not know how the error budget is allocated to each hypothesis. We formalize these two versions below. Let O be the random outcome returned by the tree.

CC-ODT with PAC-Error. An incomplete decision tree is *δ-PAC-Valid* if, for any true hypothesis h, it returns h with probability at least $1 - \delta$, formally,

$$\mathbb{P}_h(O \neq h) \leq \delta, \quad \forall h \in H.$$

CC-ODT with Total-Error. An incomplete decision tree is *δ-Total-Valid* if, for the total error probability is at most δ, formally,

$$\sum_{h \in H} \pi(h) \cdot \mathbb{P}_h[O \neq h] \leq \delta,$$

where π is the prior distribution. The goal in both versions is to find an incomplete decision tree with minimal expected cost, subject to the corresponding error constraint.

For the proof of Proposition 10, consider the PAC-error version of CC-ODT. It turns out that this version of CC-ODT is indeed quite trivial (unlike the total-error version): below we show that under PAC-error, CC-ODT is almost equivalent to the ordinary ODT problem.

Lemma 14. *Suppose $\delta \in (0, \frac{1}{2})$, and $\mathbb{T}$ is a δ-PAC-valid decision tree. Then, $\mathbb{T}$ must also be 0-valid.*

Proof. **Proof of Lemma 14** It suffices to show that there is no incomplete node in $\mathbb{T}$. For the sake of contradiction, assume $\mathbb{T}$ has an incomplete node ℓ with state (A_ℓ, p_ℓ). By the definition of incomplete node, $|A_\ell| \geq 2$, so there is an $h \in A_\ell$ with $p_\ell(h) \leq \frac{1}{2}$. Now suppose h is the true hypothesis. Since each hypothesis traces a unique path in any decision tree, regardless of whether or not it is incomplete, h will reach node ℓ with probability 1. Then at ℓ, the decision tree returns h with probability $p_\ell(h) = 1 - \sum_{g \in A_\ell : g \neq h} p_\ell(g) \leq \frac{1}{2}$, and hence $\mathbb{P}_h[O \neq h] \geq \frac{1}{2}$, reaching a contradiction. ■ □

For the reader's convenience, we recall that an ASHT instance $\mathcal{I}$ is associated with an ODT instance $\mathcal{I}_{ODT}$, defined as follows. Each action corresponds to a test $T_a : H \longrightarrow \Omega_a$ with $T_a(h) = \mu(h, a)$, where $\Omega_a = \{\mu(h, a) : h \in H\}$, and the cost T_a is $c(a) = \lceil s(a)^{-1} \log(|H|/\delta) \rceil$. Denote ODT^*_δ the minimal cost of any δ-PAC-valid decision tree for $\mathcal{I}_{ODT}$. Then we immediately obtain the following from the Lemma 14.

Corollary 3. *If $\delta \in (0, \frac{1}{2})$, then $ODT^*_0 = ODT^*_\delta$.*

Now we are able to complete the proof of the main proposition.

Proof of Proposition 10. Given a δ-PAC-error algorithm $\mathbb{A}$, we show how to construct a δ-PAC-valid decision tree $\mathbb{T}$ as follows. View $\mathbb{A}$ as a decision tree (discretize the outcome space if it is continuous). Replace each action a in $\mathbb{A}$ with the test T_a. Note that the cost of T_a is $s(a)^{-1} \log(|H|/\delta) \leq s^{-1} \log(|H|/\delta)$. Therefore by Lemma 14,

$$ODT^*_0 = ODT^*_\delta \leq c(\mathbb{T}) \leq s^{-1} \log \frac{|H|}{\delta} \cdot OPT^{FA}_\delta. \qquad ■$$

4.11 Total Error Version

In the last section we defined the total-error version of the CC-ODT problem. The total error version of the ASHT problem can be defined analogously, so we do not repeat it here. We say an algorithm is said to be δ-**total-error** if the total probability (averaged with respect to the prior π) of erroneously identified a wrong hypothesis is at most δ. The following is our main result for the total-error version.

Theorem 13. *Given an s-separated instance with uniform prior π and any $\delta \in (0, \frac{1}{4})$, for both the partially and fully adaptive versions, there exist polynomial-time δ-total-error algorithms with expected cost $O\left(s^{-1}\left(1 + |H|\delta^2\right) \log\left(|H|\delta^{-1}\right) \log |H|\right)$ times the optimum.*

In particular, if $\delta \leq O(|H|^{-1/2})$, then the above is polylog-approximation for fixed s.

We will first prove Theorem 13 for the fully adaptive version, and then show how the same proof works for the partially adaptive version. Unlike the PAC-error version where CC-ODT is almost equivalent to ODT, in the total-error version their optima can differ by a $\Omega(|H|)$ factor. We construct a sequence of ODT instances $\mathcal{I}_n$, where $n \in \mathcal{Z}^+$, with $ODT^*_\delta(\mathcal{I}_n)/ODT^*_0(\mathcal{I}_n) = O(\frac{1}{n})$.

Suppose there are $n+2$ hypotheses $h_1, ..., h_n$ and g, h, with $\pi(g) = \pi(h) = 0.49$ and $\pi(h_i) = \frac{1}{50n}$ for $i = 1, ..., 50$. Each (binary) test partitions $[n+2]$ into a singleton and its complement. Consider error budget $\delta = \frac{1}{4}$, then for each n we have $ODT^*_\delta(\mathcal{I}_n) = 1$. In fact, we may simply perform a test to separate g and h, and then return the one (out of g and h) that is consistent with the outcome. The total error of this algorithm is $1/50 < \delta$. On the other hand, $ODT^*_0(\mathcal{I}_n) = n + 1$.

However, for uniform prior, this gap is bounded:

Proposition 15. *Suppose the prior π is uniform. Then, for any $\delta \in (0, \frac{1}{4})$, it holds*

$$ODT^*_0 \le \left(1 + O(|H|\delta^2)\right) \cdot ODT^*_\delta.$$

To show the above, we need the following building block.

Lemma 15. *Suppose the prior π is uniform. Then, for any $\delta \in [0, \frac{1}{4})$, the total prior probability density on the incomplete nodes is bounded by $\sum_{\ell \text{ inc.}} \pi(A_\ell) \le 2\delta$.*

Proof. **Proof of Lemma 15** Let ℓ be an incomplete node with state (A_ℓ, p_ℓ) and write $p = p_\ell$ for simplicity. Then, the error probability contributed by ℓ is

$$\begin{aligned}\sum_{h \in A_\ell} \pi(h) \cdot (1 - p(h)) &= \sum_{h \in A_\ell} \pi(h) - \sum_{h \in A_\ell} \pi(h) \cdot p(h) \\ &= \pi(A_\ell) - \frac{1}{n} \sum_{h \in A_\ell} p(h) \\ &= \frac{|A_\ell|}{n} - \frac{1}{n} \ge \frac{1}{2}\pi(A_\ell),\end{aligned}$$

where the last inequality follows since $|A_\ell| \ge 2$. By the definition of δ-PAC-error, it follows that

$$\delta \ge \sum_{\ell \text{ inc.}} \sum_{h \in A_\ell} \pi(h) \cdot (1 - p(h)) \ge \frac{1}{2} \sum_{\ell \text{ inc.}} \pi(A_\ell),$$

i.e., $\sum_{\ell \text{ inc.}} \pi(A_\ell) \le 2\delta$. ■ □

Proof of Proposition 15. It suffices to show how to convert a decision tree $\mathbb{T}$ with δ-total-error to one with 0-total-error, without increasing the cost by too much. Consider each incomplete node A_ℓ in $\mathbb{T}$. We will replace A_ℓ with a (small) decision tree that uniquely identifies a hypothesis in A_ℓ. Consider any distinct hypotheses $g, h \in A_\ell$. Then by Assumption 3, there is an action $a \in A$ with $d(g, h; a) \ge s$. So if we select T_a, then by Hoeffding bound (Theorem 1), we have that with high probability at least one of g and h will be eliminated, and the number of alive hypotheses in A_ℓ reduces by at least 1. Thus, by repeating this procedure iteratively for at most $|A_\ell| - 1$ times, we can identify a unique hypothesis. Since each test T_a corresponds to selecting a

for $c(a) = s(a)^{-1}\log(|H|/\delta) \le s^{-1}\log(|H|/\delta)$ times in a row, this procedure increases the total cost by $\sum_{\ell \text{ inc.}} \pi(A_\ell) \cdot (|A_\ell| \cdot s^{-1}\log(|H|/\delta)$. Therefore,

$$\begin{aligned} ODT_0^* &\le ODT_\delta^* + \sum_{\ell \text{ inc.}} \pi(A_\ell)|A_\ell| s^{-1} \log\frac{|H|}{\delta} \\ &= ODT_\delta^* + \sum_{\ell \text{ inc.}} \pi(A_\ell)|H|\pi(A_\ell) \cdot s^{-1}\log\frac{|H|}{\delta} \\ &= ODT_\delta^* + O\Big(s^{-1}|H|\log\frac{|H|}{\delta} \cdot \sum_{\ell \text{ inc.}} \pi(A_\ell)^2\Big). \end{aligned} \tag{4.6}$$

Since $\sum_{\ell \text{ inc.}} \pi(A_\ell) \le 2\delta$ and each $\pi(A_\ell)$'s is non-negative, we have $\sum_{\ell \text{ inc.}} \pi(A_\ell)^2 \le \left(\sum_{\ell \text{ inc.}} \pi(A_\ell)\right)^2 \le 4\delta^2$. Further, by Pinsker's inequality, we have $ODT_\delta^* = \Omega(s^{-1}\log\frac{|H|}{\delta})$. Combining these two facts with Eq. (4.6), we obtain $ODT_0^* \le \big(1 + O(|H|\delta^2)\big) \cdot ODT_\delta^*$. ■

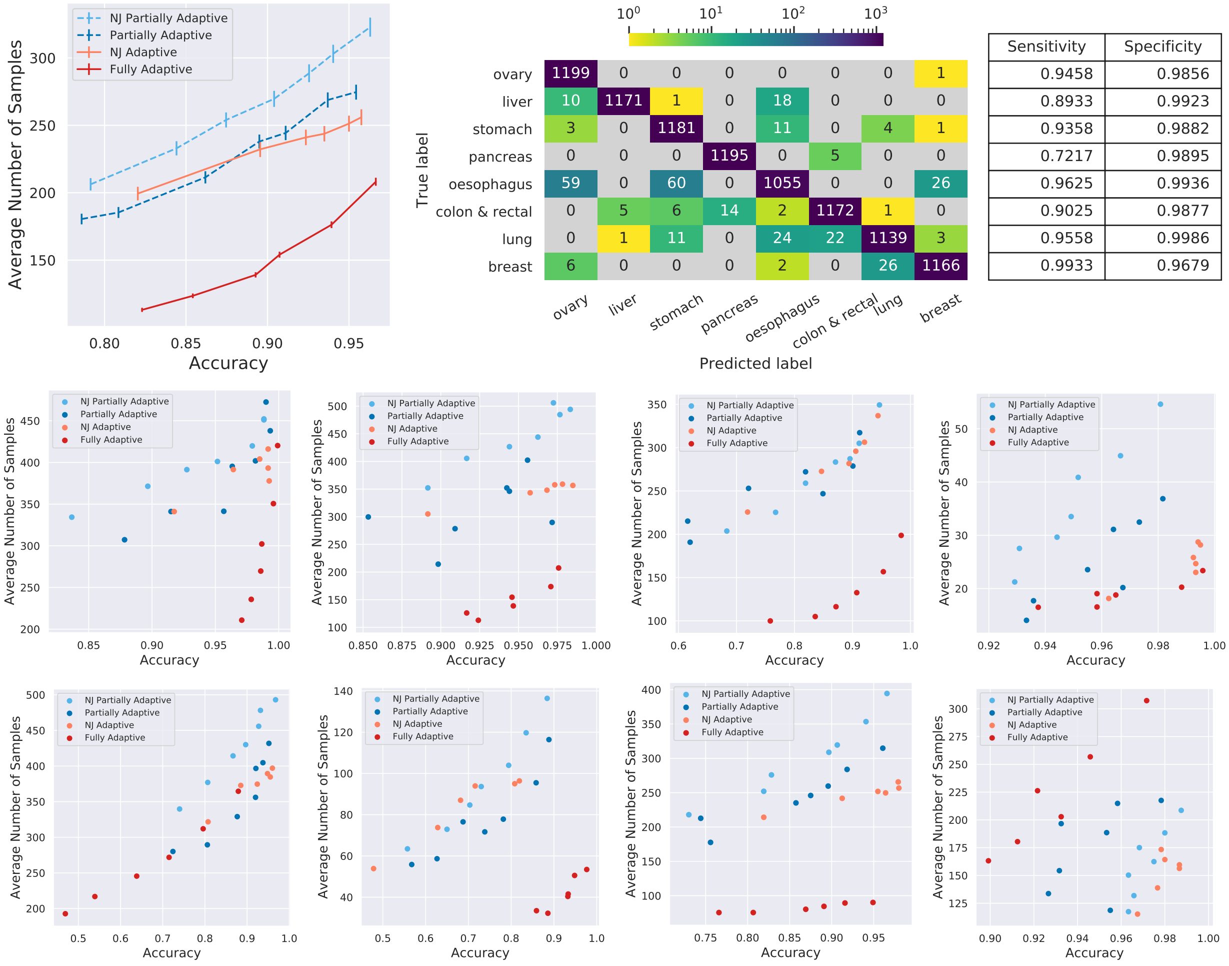
The following lemma can be proved using the same idea of the proof of Proposition 10.

Lemma 16. $ODT_\delta^* \le O\big(s^{-1}\log(|H|/\delta)\big) OPT_\delta^{FA}$.

Now we are ready to show Theorem 13.

$$\begin{aligned} GRE &\le O(\log|H|) \cdot ODT_0^* && \text{(Theorem 15)} \\ &\le O\big((1 + O(|H|\delta^2)\big)\log|H|) \cdot ODT_\delta^* && \text{(Lemma 15)} \\ &\le O\big((1 + O(|H|\delta^2)\big) s^{-1}\log^2\frac{|H|}{\delta}\log|H|) \cdot OPT_\delta^{FA}. && \text{(Lemma 16)} \end{aligned}$$

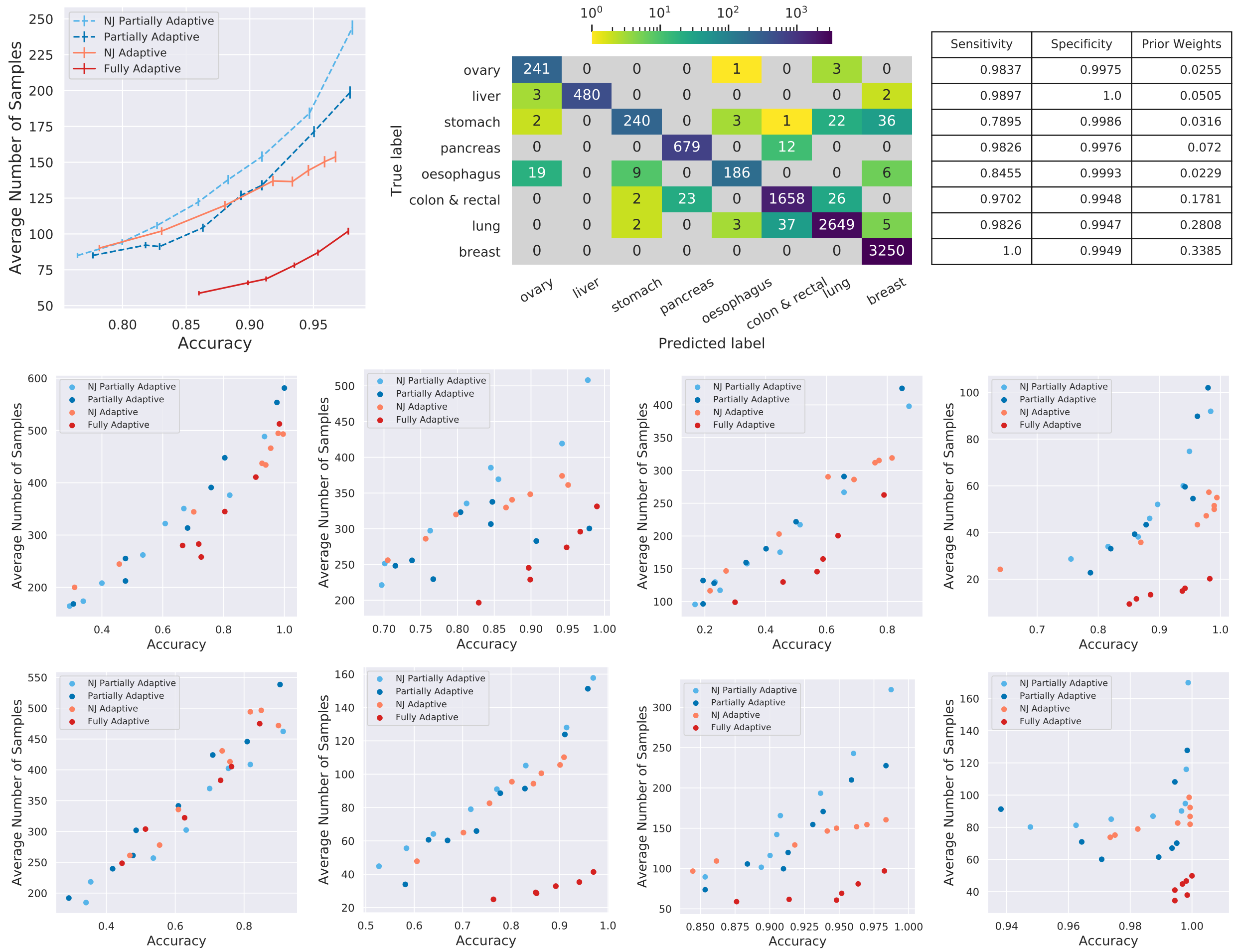
The above proof can be adapted to the partially adaptive version straightforwardly as follows. Observing that partially adaptive algorithms can be viewed as a special case of the fully adaptive, we can define $ODT_{0,\mathrm{PA}}^*$ and $ODT_{\delta,\mathrm{PA}}^*$ (analogous to ODT_0^* and ODT_δ^*) for the partially adaptive version, as the optimal cost of any partially adaptive decision tree with 0 or δ error. By replacing ODT_δ^* and ODT_0^* with $ODT_{0,\mathrm{PA}}^*$ and $ODT_{\delta,\mathrm{PA}}^*$, one may immediatly verify that inequalities in Lemmas 15 and 16 hold for the partially adaptive version. Furthermore, the first inequality above can be established for the partially adaptive version by replacing Theorem 15 with Theorem 14, hence completing the proof.



True label \ Predicted label	ovary	liver	stomach	pancreas	oesophagus	colon & rectal	lung	breast
ovary	1199	0	0	0	0	0	0	1
liver	10	1171	1	0	18	0	0	0
stomach	3	0	1181	0	11	0	4	1
pancreas	0	0	0	1195	0	5	0	0
oesophagus	59	0	60	0	1055	0	0	26
colon & rectal	0	5	6	14	2	1172	1	0
lung	0	1	11	0	24	22	1139	3
breast	6	0	0	0	2	0	26	1166

Sensitivity	Specificity
0.9458	0.9856
0.8933	0.9923
0.9358	0.9882
0.7217	0.9895
0.9625	0.9936
0.9025	0.9877
0.9558	0.9986
0.9933	0.9679

Figure 4.3: Comparison of our partially and fully adaptive algorithms with those of NJ's on real-world data, COSMIC, under *uniform prior*. Top row (left): each point is averaged over 9,600 replications. The error bars are the 95 percentage confidence intervals for the estimated means. Top row (middle): the confusion matrix of *Fully Adaptive* where the algorithm accuracy equals to 0.97, and each row sums up to 1,200. Top row (right): the sensitivity and specificity our algorithm (top row middle) for each cancer type. Middle and Bottom rows: the sensitivity figures for each cancer type. The order of the figures follows the order that they appear the confusion matrix (top row right). Each point in these figures is averaged over 1,200 replications contained in the top left figure.



	ovary	liver	stomach	pancreas	oesophagus	colon & rectal	lung	breast
ovary	241	0	0	0	1	0	3	0
liver	3	480	0	0	0	0	0	2
stomach	2	0	240	0	3	1	22	36
pancreas	0	0	0	679	0	12	0	0
oesophagus	19	0	9	0	186	0	0	6
colon & rectal	0	0	2	23	0	1658	26	0
lung	0	0	2	0	3	37	2649	5
breast	0	0	0	0	0	0	0	3250

Sensitivity	Specificity	Prior Weights
0.9837	0.9975	0.0255
0.9897	1.0	0.0505
0.7895	0.9986	0.0316
0.9826	0.9976	0.072
0.8455	0.9993	0.0229
0.9702	0.9948	0.1781
0.9826	0.9947	0.2808
1.0	0.9949	0.3385

Figure 4.4: Comparison of our partially and fully adaptive algorithms with those of NJ's on real-world data, COSMIC, under non-uniform prior. Top row (left): each point is averaged over 9,600 replications. The error bars are the 95 percentage confidence intervals for the estimated means. Top row (middle): the confusion matrix of *Fully Adaptive* where the algorithm accuracy equals to 0.97. The sum of each row (the number of replications for each cancer type) equals to the corresponding prior multiplied by 9,600 and rounded to the nearest integer. Top row (right): the sensitivity and specificity our algorithm (top row middle) for each cancer type. Middle and Bottom rows: the sensitivity figures for each cancer type. The order of the figures follows the order that they appear the confusion matrix (top row right). The number of replications that we average over for each point in each figure equals to the number of replications for each cancer type.

Chapter 5

Machine Learning Algorithms for Predicting Hospital Readmissions in Sickle Cell Disease

5.1 Introduction

Sickle Cell Disease (SCD) is the most common inherited hemoglobinopathy worldwide and carries high morbidity and mortality Maitra et al. (2017), Mehari et al. (2012). Complications related to SCD have resulted in prolonged hospitalizations and high frequency of 30-day hospital readmissions Benenson et al. (2017), AlJuburi and Majeed (2013), Brodsky et al. (2017), Brousseau et al. (2010), Joynt et al. (2012), Machado et al. (2011), Nouraie and Gordeuk (2015). For example, in the largest retrospective multi-state study of 21,112 adult patients with SCD in the United States, 33.4% of patients had 30-day readmission with 22.1% readmitted within 14 days Brousseau et al. (2010). Other studies found that 50% of adult patients with SCD were readmitted within 30 days, and those who returned within one week had the poorest overall prognosis Frei-Jones et al. (2009), Ballas and Lusardi (2005). As policymakers are mandating the implementation of evidence-based quality improvement interventions, the frequency of 30-day hospital readmissions becomes an important clinical metric to assess the quality of care amongst chronic diseases, including SCD Wilson-Frederick et al. (2019). Hospital readmission risk has been traditionally calculated using simple scoring systems (such as the LACE and HOSPITAL indices) with limited features van Walraven et al. (2010), Donzé et al. (2013), and not specific to high-risk groups such as patients with SCD, where socio-economic factors may play an important role in hospital readmissions Kansagara et al. (2011), Cronin et al. (2019), Adzika et al. (2017), Brown et al. (2015), Chen et al. (2019). For instance, the LACE index was validated on a Canadian middle-age population with

ICD-9	282.41, 282.42, 282.6, 282.60, 282.61, 282.62, 282.63, 282.64, 282.68, 282.69
ICD-10	D57.0, D57.00, D57.01, D57.02, D57.1, D57.2, D57.20, D57.21, D57.211, D57.212, D57.219, D57.3, D57.4, D57.40, D57.41, D57.411, D57.412, D57.419, D57.8, D57.80, D57.81, D57.811, D57.812, D57.819

Table 5.1: Sickle cell ICD-9 ICD-10 diagnosis codes. ICD-10 D57.3 (sickle cell trait) was removed from the inclusion criteria. After removing patients who were only diagnosed with sickle cell trait (ICD-10 D57.3) during the study period, we had 1009 patients left in the dataset.

few comorbidities van Walraven et al. (2010), and therefore it does not capture the demographics and disease-specific complexities that are inherent in the SCD population. In fact, the predictors of hospital readmission in patients with SCD are currently not being evaluated in clinical practice. One limitation of standard models to predict hospital readmissions is that they are hypothesis-driven; they use a fixed set of predictive features and may ignore disease-specific features that can impact clinical outcomes. Machine Learning (ML) algorithms–a class of algorithms that can be used in detecting underlying patterns in high dimensional datasets–can potentially be a useful tool in predicting hospital readmission risks in the SCD patient population. In many healthcare applications, the performance of ML algorithms has dominated that of traditional statistical methods Chen et al. (2017), Hsich et al. (2011), Gorodeski et al. (2011), Chen et al. (2011), Amalakuhan et al. (2012), Chirikov et al. (2017), Thottakkara et al. (2016), and several studies have employed ML algorithms to predict 30-day hospital readmissions Mortazavi et al. (2016), Xue et al. (2018), Weinreich et al. (2016), Shameer et al. (2017), Eckert et al. (2019), Deschepper et al. (2019), Futoma et al. (2015). However, none of them has been conducted on the high-risk SCD patient population. The objective of this research is to explore the value of ML algorithms, combined with domain knowledge, in predicting hospital readmission risk for a SCD patient population using a real-world data source Sherman et al. (2016). Specifically, we used both clinical knowledge-driven and hypothesis-free data features extracted from electronic health records (EHR) data to guide our ML models. We hypothesized that ML algorithms would (a) outperform traditional risk scoring systems, (b) find a richer set of predictors that can better guild clinical practice, and hence (**C**) be a more suitable tool in predicting hospital readmission risk among the SCD patient population.

5.2 Materials and Methods

Design and Sample The University of Pittsburgh Medical Center (UPMC) Institutional Review Board approved this study. The R3 services through the Department of Bioinformatics served

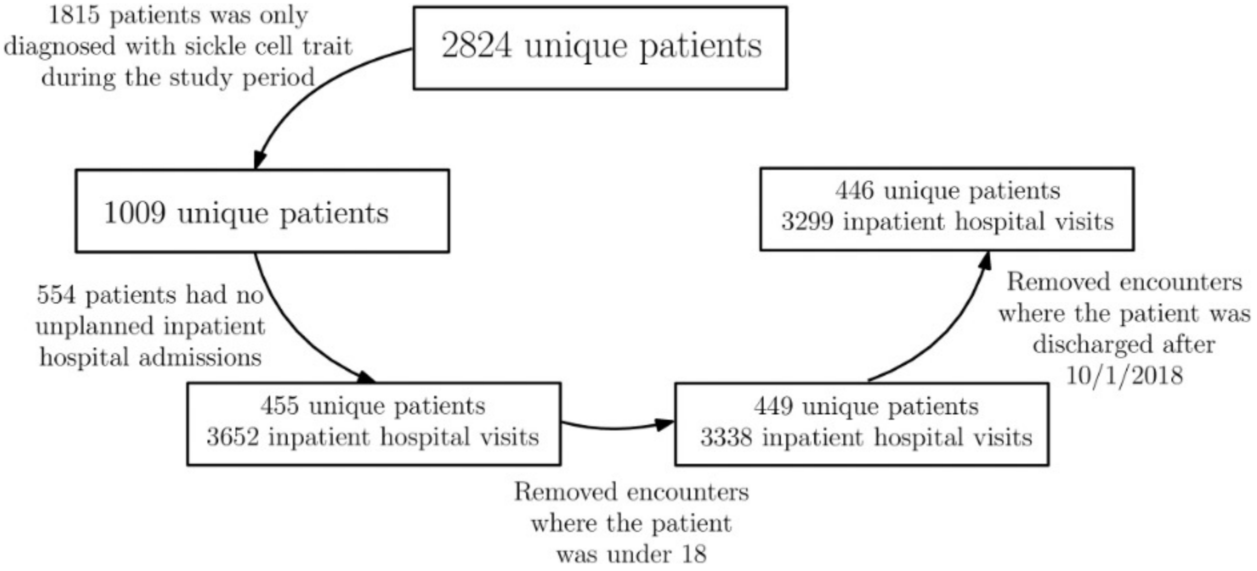

2824 unique patients
1815 patients was only diagnosed with sickle cell trait during the study period
1009 unique patients
554 patients had no unplanned inpatient hospital admissions
455 unique patients
3652 inpatient hospital visits
Removed encounters where the patient was under 18
449 unique patients
3338 inpatient hospital visits
Removed encounters where the patient was discharged after 10/1/2018
446 unique patients
3299 inpatient hospital visits

Figure 5.1: Study inclusion criteria flow chart. Description of the patient and inpatient visit inclusion criteria. 455 patients had at least one unplanned inpatient visit from January 1, 2013 to November 1, 2018. All consecutive (n=15) unplanned inpatient admissions where the discharge and readmission dates are the same were combined. We removed any inpatient encounters in which the patient was under the age of 18 at the time of the visit given that we did not have access to the Children's Hospital EHR database. Since inpatient visits after October 1, 2018 were censored, we removed those visits and resulted in 446 patients and 3299 unplanned inpatient visits.

as an honest data broker to ensure all patient health information was de-identified and Health Insurance Portability and Accountability Act-compliant throughout the research cycle, including but not limited to data extraction, data management, analytical and machine learning processes. All analyses were conducted on de-identified patient data. Our SCD patient cohort was selected from five hospitals across the UPMC hospital system, where patients with SCD are followed by the adult UPMC Sickle Cell Program's inpatient consult service. The UPMC Sickle Cell Program is the only provider of specialized care for SCD in the region, and thus only a negligible number of patients with SCD is admitted to hospitals where the UPMC Sickle Cell Program staff has no clinical privileges. The raw data contains the EHR data of 2824 patients selected by the principal diagnosis of SCD using the ICD-9 and ICD-10 codes listed in Table 5.1 Quan et al. (2005), Snyder et al. (2017) between January 1, 2013, and November 1, 2018. The preprocessed dataset contains 446 patients and 3299 unplanned inpatient visits, and Figure 5.1 summarizes the patient inclusion criteria of this study.

Outcome variables An admission was defined as an unplanned inpatient hospital admission, identified by a non-elective hospital admission type as indicated by the EHR data. A readmission was defined as an admission within 30 days of the discharge date of the last admission. We excluded any admission to a maternity unit, skilled nursing facility, and rehabilitation unit. In our study, a case was defined as an admission that resulted in a readmission, while a control was indicated by an admission that did not result in a readmission.

Predictor Candidates All analyses have been conducted on the de-identified patient dataset, and patients who were admitted to other hospitals not defined above were not captured. The preprocessed features (n = 481), including labs, demographics, the number of outpatient visits prior to the current visit, and the number of emergency department (ED) visits prior to the current visit Kansagara et al. (2011), Cronin et al. (2019), Brom et al. (2020), were extracted from the EHR data using both data-driven methods and clinical knowledge (Table 5.5 in § 5.5). The dataset also includes 21 variables extracted according to the LACE van Walraven et al. (2010) and HOSPITAL Donzé et al. (2013) indices: the length of stay, the number of ED visits in the past 6 months, the number of (unplanned) hospital admissions in the past year, whether any procedure was performed during the hospitalization, and 17 ICD-9/ICD-10 code groups to calculate the Charlson comorbidity index score in the LACE index. The remaining features included 340 ICD-9/ICD-10 diagnosis codes, 2 demographic features, 4 healthcare insurance provider types, 42 medication groups, 13 lab categories, 25 procedures, 2 zip codes, 5 smoking status features, 7 vital signs, 34 hospital departments, the number of outpatient visits (prior to the current visit in the study period). To further capture the trend in patient readmission patterns, we included additional variables: the number of ED visits (prior to the current visit) in the study period, the number of days since the last inpatient visit (of the current visit), and the number of inpatient visits (prior to the current visit) in the study period. We included labs that were processed through a centralized lab and eliminated point of care testing.

Data Preprocessing Each lab variable takes 6 categorical values (Tables 5.5 and 5.6 in § 5.5) to indicate whether a lab result is missing, normal, low, high, low panic, or high panic. All lab variables were defined based on central lab reference values and were not adjusted to the normalized lab values for an individual patient. The vital sign variables were kept as continuous in the RF model and were preprocessed into categorical variables in the LR and SVM models. Table 5.5 (in § 5.5) includes the cutoff values for preprocessing these variables into categorical variables. The reason why the vital sign variables were coded as continuous instead of predefining the cutoff values using domain knowledge as in the RF model is that the RF algorithm automatically selects cutoff values that have high predictive value (indeed, this is one of the RF algorithm's advantages). Out of the 3299 encounters, 873 (26.5%) did not have any vital signs taken; 685 (20.%) did not have smoking status; 656 (19.8%) did not have any medication prescriptions; 454 (13.8%) did not have any procedures performed; 39 (1.2%) did not have any lab tests. The latter three could be classified as missing values or not applicable depending on the individual patient circumstance. The rest of the data does not contain any missing information. Table 5.5 (in § 5.5) describes the percentage available information for each individual variable in detail. Instead of imputing the missing values, we created a dummy variable for each variable that contains missing information to indicate whether this variable is missing in a particular encounter. This is a popular method in the ML community to handle missing data, and has shown superiority to other methods in healthcare applications where data is not missing at random but rather a reflection of the decision made by care providers Marlin et al. (2012), Lipton et al. (2016). Twenty-seven out of 195 readmission patients died during the observation period, and these 27 patients had 200 unplanned inpatient admissions in total. Sixty out of 446 total patients died during the observation period, and these 60 patients had 247 unplanned inpatient admissions. Since the number of admissions that resulted in mortality was less than 2%, those admissions were kept in the training and testing dataset.

Methods To predict whether an inpatient visit will result in a readmission, three standard ML algorithms were applied using the scikit-learn package in Python: Logistic Regression (LR) Kleinbaum et al. (2002), Support Vector Machine (SVM) Cortes and Vapnik (1995), and Random Forest (RF) Breiman (2001). Traditional risk scoring systems, the LACE and HOSPITAL indices, were also applied van Walraven et al. (2010), Donzé et al. (2013). Although LACE and HOSPITAL have not been previously applied to the SCD patient population, they provide two benchmark models for comparison. All variables needed to compute those two indices were contained in the EHR data. In addition, to test the impact of patients with frequent admissions on our ML models, we included a weighted RF model where each admission is weighted inversely by the total number of admissions incurred by the patient during the study period. § 5.6 describes the details of each algorithm and how they were used. The features mentioned above were treated as inputs to these models. We randomly selected the admissions incurred by 30% of the 195 return patients and 251 non-return patients to be the testing set (n = 134); the training set contained the admissions incurred by the remaining 211 patients. Thus, our training and testing sets contained the same demographic information, predictors, and outcomes.

Model Evaluation We used the C-statistic, or equivalently the Area Under the Receiver Operating Characteristic Curve (AUC), and precision-recall curves as two quantitative metrics for identifying predictive performance within each of the classifiers. For intuition: a perfect classifier achieves a C-statistic of 1, while random chance corresponds to a C-statistic of 0.5. In addition, we reported the sensitivity and specificity of our best performing model. Since the number of samples in our study was relatively small, our results might have been sensitive to different training and

		Total N = 446	Readmission Group N = 195
General	Unplanned inpatient encounters	3299	2823*
	Number of ED visits	6780	4899
	Number of outpatient visits	10731	5978
	Average length of stay per admission	5.895 (5.974)	5.543 (5.586)
	Average number of admissions per patient	7.40 (12.90)	14.47 (16.97)
	Number with HbSS	255	139
	Number with HbSC	55	30
	Number with HbS/B0 or HbS/B+	36	26
Age	18-29	157	80
	30-49	136	56
	50-69	108	44
	70-89	42	14
	≥ 90	3	1
Gender	Male	175	70
	Female	271	125
LACE Index		10.26 (2.79)	10.52 (2.73)
HOSPITAL Index		8.16 (2.40)	8.53 (2.32)

Table 5.2: Characteristics and Demographics of the Post-Processed Data Set. Description of the 3299 encounters of the 446 patients included in the post-processed data set. We also included the distribution of LACE and HOSPITAL indices computed using the EHR data. The sickle cell genotypes HbS/B0 and HbS/B+ are grouped into one genotype since ICD-9 diagnosis codes do not distinguish between these two genotypes. *: 1369 (out of 2823) were readmissions.

testing splits. To address this problem, we performed 100 different splits and averaged the resulting 100 C-statistics.

5.3 Study Results

Our training and testing sets contained the same demographic information, predictors and outcomes. Table 5.2 summarizes the characteristics and demographics of the post-processed data set, as well as the distributions of LACE and HOSPITAL indices computed using the post-processed data. Our cohort included 3299 admissions of 446 adult patients with SCD. Of these patients, 195

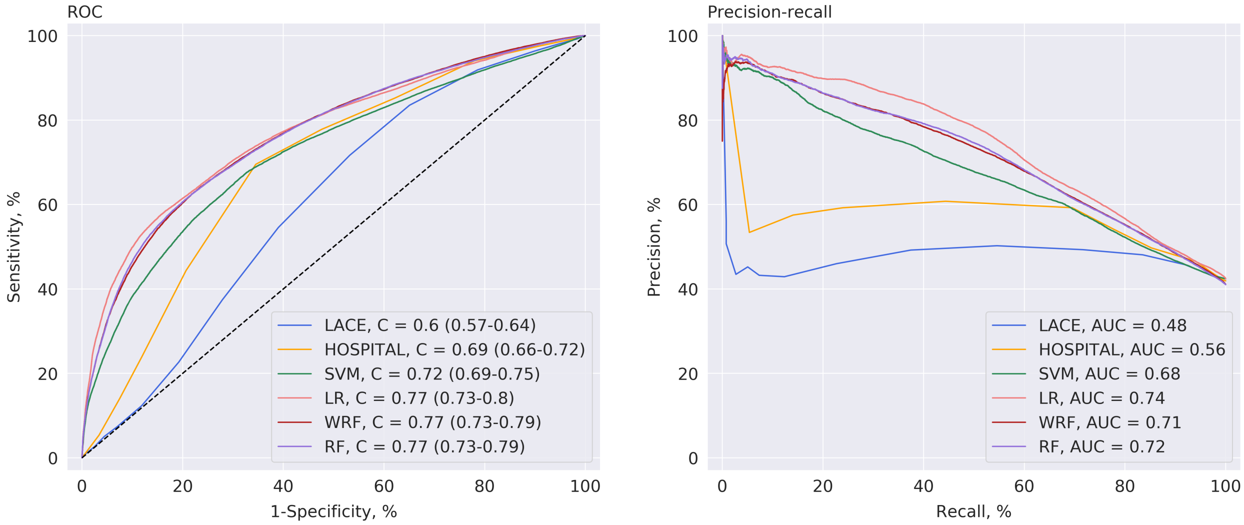


Figure 5.2: Performance Metrics of Machine Learning Models for Predicting 30-Day Readmissions in Sickle Cell Disease. Two performance metrics measured out-of-sample and averaged over 100 independent train/test draws. (A) Receiver operating characteristic curves, and corresponding area under the curve; also known as the C-statistic. (B) Precision-recall curves.

(43.72% of readmission) patients were readmitted within 30 days for a total of 1369 times. The average age of those 446 patients was 42.22 (SD = 19.03) years, and the average age of the 195 patients who had readmission during the study period was 39.47 (18.14) years. The average LACE and HOSPITAL indices of those 3299 admission were 10.26 (2.79) and 8.16 (2.40), respectively.

To prevent overfitting, in the LR model we added LASSO regularization, and in the RF model we restricted the maximum depth of the decision trees to 15. Fig. 5.2 summarizes the two performance metrics of each model–the Receiver Operating Characteristic (ROC) and precision-recall curves. LACE had a C-statistic of 0.6 (95%CI 0.57-0.64); HOSPITAL performed slightly better than LACE (C-statistic 0.69, 95%CI 0.66-0.72); SVM with 'rbf' kernel outperformed HOSPITAL in terms of C-statistic (C-statistic 0.72, 95%CI 0.69-0.75); LR outperformed SVM by a large margin (C-statistic 0.77, 95%CI 0.73-0.8); RF performed similar to logistic regression (C-statistic 0.77, 95%CI 0.73-0.79). Furthermore, the weighted random forest (C-statistic 0.77, 95%CI 0.73-0.79) model performs similar to the random forest model. Similarly, in terms of precision-recall, SVM (Area Under the Curve (AUC) 0.68) outperformed HOSPITAL (AUC 0.56), and the RF model (AUC 0.74) and the LR model (AUC 0.72) performed the best. In both the ROC and precision-recall curves, we observed that the curves corresponding to RF and LR pointwise dominate these of LACE and HOSPITAL indices.

Having established that the RF and LR models had the best performance, we compare the sensitivities and specificities of those two models against these of the LACE and HOSPITAL indices in Tables 5.3 and 5.4 respectively. In Tables 5.3 and 5.4, the thresholds were chosen to

Model		Predicted Positive (%)	Predicted Negative (%)		
RF	True Positive (%)	39.61	19.41	Sensitivity (%)	67.1 ± 3.8
	True Negative (%)	11.62	29.37	Specificity (%)	71.1 ± 4.3
LR	True Positive (%)	39.42	18.20	Sensitivity (%)	68.4 ± 3.8
	True Negative (%)	12.02	30.36	Specificity (%)	71.1 ± 4.3
LACE	True Positive (%)	27.19	30.89	Sensitivity (%)	46.8 ± 4.1
	True Negative (%)	11.89	30.04	Specificity (%)	71.7 ± 4.3

Table 5.3: Out-of-sample Prediction Performance of the Random Forest and Logistic Regression Models Compared to LACE Index. Confusion matrices and corresponding sensitivities and specificities for the random forest and logistic regression classifier. A true positive (negative) case was determined as the admission did (not) result in a 30-day readmission and we correctly predicted so. The threshold of the LACE index is chosen to be 10 (van Walraven et al. 2010). The thresholds of RF and LR are chosen such that the specificities of these models match the specificity of the LACE index. Results are averaged over 100 independent train/test draws, where an average test set contains 134 patients and 1000 visits. Sensitivity and specificity are reported with 95% confidence intervals.

match the specificities of RF and LR models to those of the LACE index and HOSPITAL index, respectively. A true negative case was determined as a hospital admission that did not result in a 30-day readmission, and we correctly predicted so, and a true positive case was determined as a hospital admission that did result in a 30-day readmission, and we also correctly predicted so. In Tables 5.3 and 5.4, we again observed that the performances of RF and LR were similar in terms of sensitivity at their corresponding chosen thresholds, and the sensitivities of both models outperformed those of the LACE index and HOSPITAL index, respectively.

To check the clinical integrity of our models, we reported the selected a subset of variables and reported their importance factors in our RF model (Deschepper et al. 2019) (see Fig. 5.3) and in our LR model (Fig. 5.4). While the variables below the selected important predictors from the LR model have near 0 coefficients (i.e., they have minimal impact on the prediction outcome), the variables outside the selected important predictions from the RF model could still have relatively large impacts on the model. Thus, in Fig. 5.3 we provided the average information gain (the amount of improvement in classification) of the selected variables appearing in the random forest model, and in Fig. 5.4, we reported both the direction and the standardized magnitude of coefficients of the selected variables in the LR model. Both Fig.s 5.3 and 5.4 contain similar features.

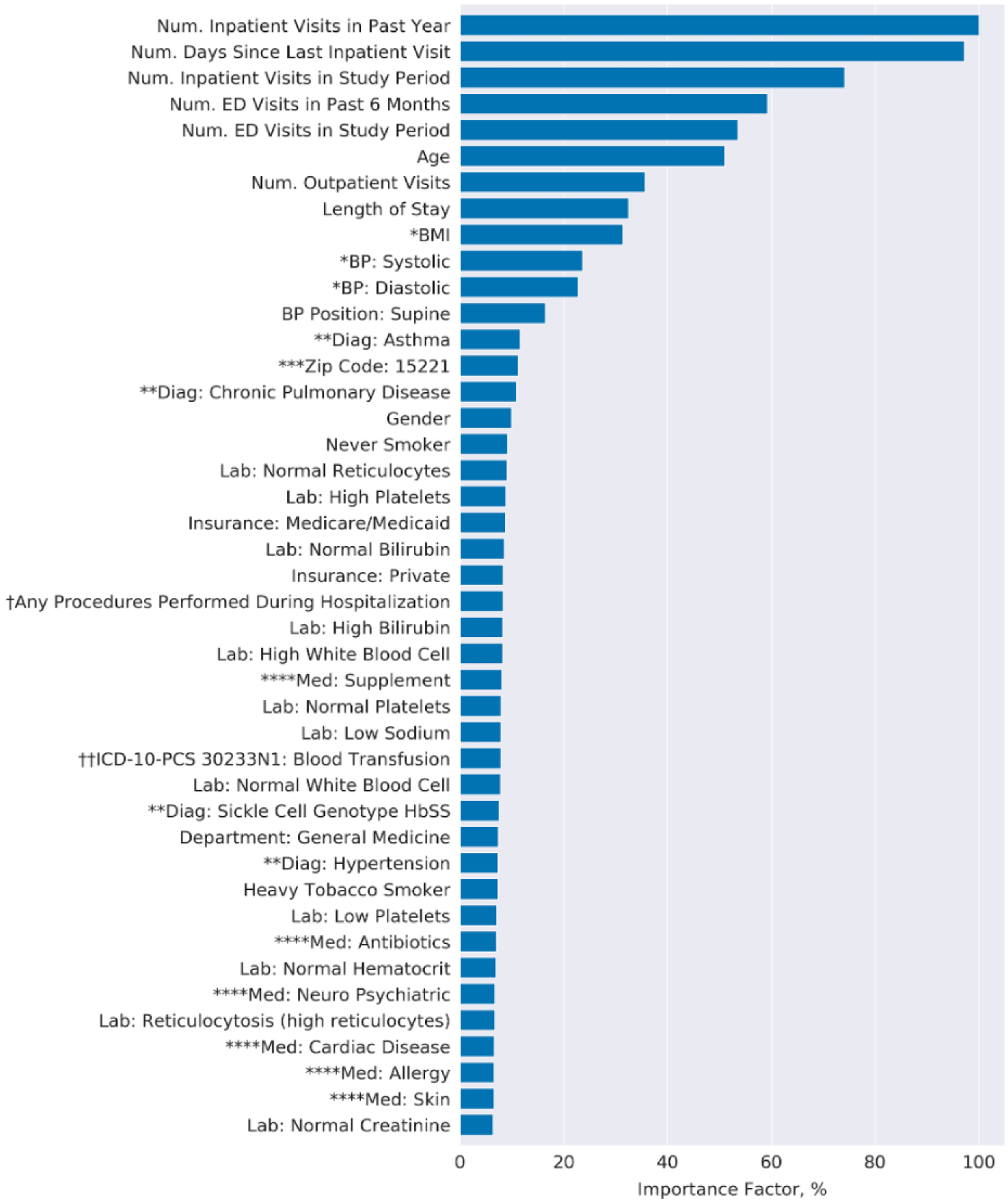

Num. Inpatient Visits in Past Year
Num. Days Since Last Inpatient Visit
Num. Inpatient Visits in Study Period
Num. ED Visits in Past 6 Months
Num. ED Visits in Study Period
Age
Num. Outpatient Visits
Length of Stay
*BMI
*BP: Systolic
*BP: Diastolic
BP Position: Supine
**Diag: Asthma
***Zip Code: 15221
**Diag: Chronic Pulmonary Disease
Gender
Never Smoker
Lab: Normal Reticulocytes
Lab: High Platelets
Insurance: Medicare/Medicaid
Lab: Normal Bilirubin
Insurance: Private
†Any Procedures Performed During Hospitalization
Lab: High Bilirubin
Lab: High White Blood Cell
****Med: Supplement
Lab: Normal Platelets
Lab: Low Sodium
††ICD-10-PCS 30233N1: Blood Transfusion
Lab: Normal White Blood Cell
**Diag: Sickle Cell Genotype HbSS
Department: General Medicine
**Diag: Hypertension
Heavy Tobacco Smoker
Lab: Low Platelets
****Med: Antibiotics
Lab: Normal Hematocrit
****Med: Neuro Psychiatric
Lab: Reticulocytosis (high reticulocytes)
****Med: Cardiac Disease
****Med: Allergy
****Med: Skin
Lab: Normal Creatinine
0
20
40
60
80
100
Importance Factor, %

Figure 5.3 *(previous page)*: Important Predictors for 30-Day Readmissions in Sickle Cell Disease Selected by Random Forest Model. Importance scores of a subset of the most important variables selected by the random forest model, averaged over the 100-independent train/test draws. Importance is a measure of each variable's cumulative contribution toward reducing square error, or heterogeneity within the subset, after the data set is sequentially split according to that variable. Thus, importance reflects a variable's significance in prediction. Absolute importance is then scaled to give relative importance, with a maximum importance of 100. Since the decision boundary of the random forest is extremely non-linear, the features above are not associated with directions. Although the random forest model is less interpretable, it can model more complex relations between variables. *The vital signs in the RF model are continuous as explained in the Data Preprocessing Section. **Diag: Asthma corresponds to the ICD-9 codes that start with 493. and the ICD-10 codes J44.0-J45; Diag: Chronic Pulmonary Disease corresponds to the following ICD-9/ICD-10 codes: 416.8, 416.9, 490.-505., 506.4, 508.1, 508.8, I27.8, I27.9, J40.-J47., J60.-J67., J68.4, J70.1, J70.3; Diag: Sickle Cell Genotype HbSS corresponds to the following ICD-9/ICD-10 codes: 282.62, 282.61, D57.0, D57.00, D57.01, D57.02; Diag: Hypertension corresponds to the following ICD-9/ICD-10 codes: 401-405, I16., I10.-I13., I15., N26.2. ***Zip code 15221 corresponds to the borough of Wilkinsburg, PA, within the Pittsburgh metropolitan area. †Any procedures performed during the hospitalization is one of variables included by the LACE and HOSPITAL indices. ††ICD-10-PCS procedure code 30233N1 corresponds to "transfusion of nonautologous red blood cells into peripheral vein, percutaneous approach." ****Med: Supplement includes all dietary supplements; Med: Infection indicated whether a patient was prescribed with any antibiotics during the hospitalization (this variable is used to indicate whether the patient has any bacteria infection in addition to the ICD-9/ICD-10 coding); similarly, Med: Neuro Psychiatric includes all antipsychotic medications; Med: Cardiac Disease includes all cardiac medications; Med: Allergy and Med: Skin include all medications that can be used to treat allergy and skin problems, respectively.

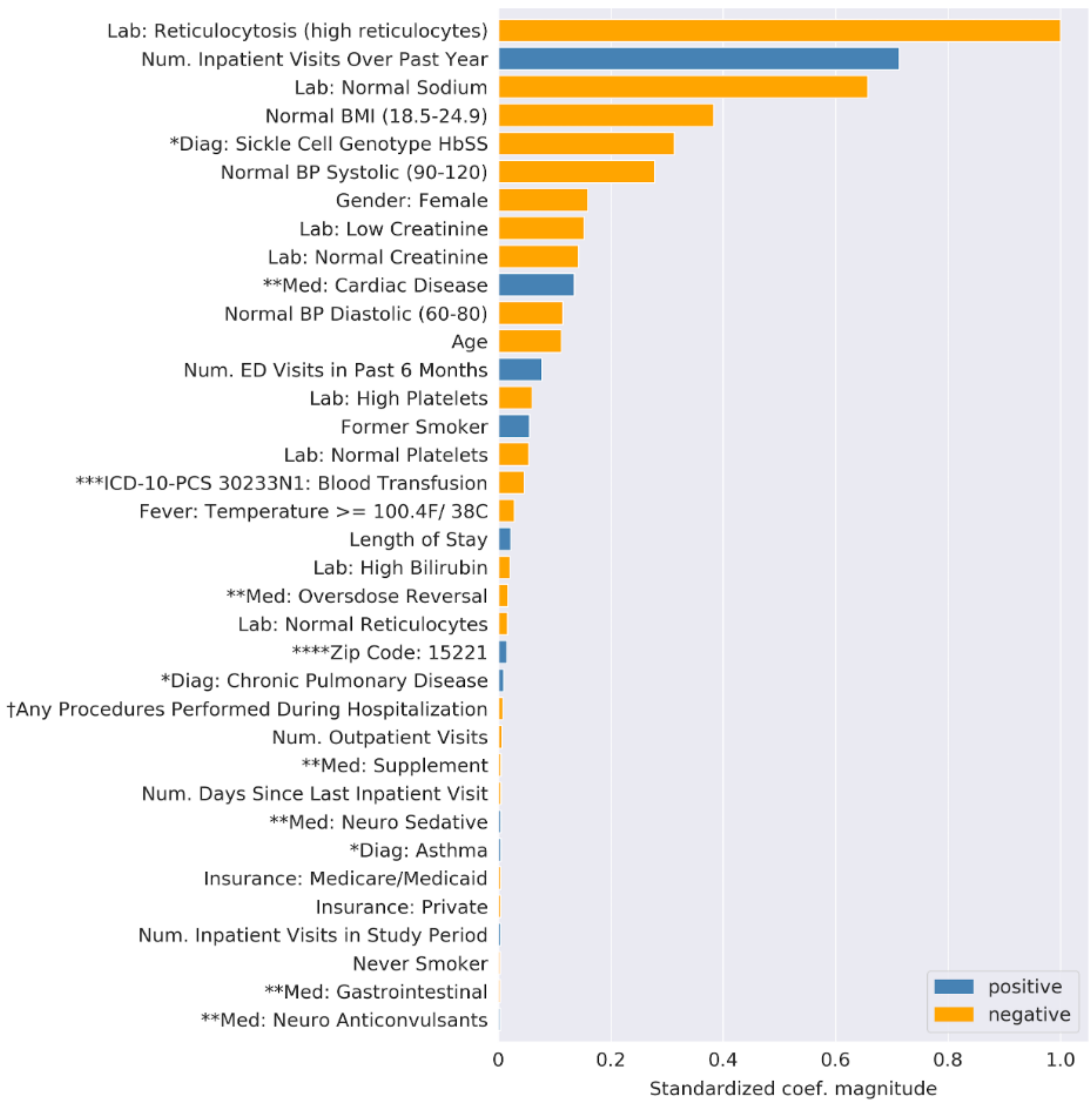

Lab: Reticulocytosis (high reticulocytes)
Num. Inpatient Visits Over Past Year
Lab: Normal Sodium
Normal BMI (18.5-24.9)
*Diag: Sickle Cell Genotype HbSS
Normal BP Systolic (90-120)
Gender: Female
Lab: Low Creatinine
Lab: Normal Creatinine
**Med: Cardiac Disease
Normal BP Diastolic (60-80)
Age
Num. ED Visits in Past 6 Months
Lab: High Platelets
Former Smoker
Lab: Normal Platelets
***ICD-10-PCS 30233N1: Blood Transfusion
Fever: Temperature >= 100.4F/ 38C
Length of Stay
Lab: High Bilirubin
**Med: Oversdose Reversal
Lab: Normal Reticulocytes
****Zip Code: 15221
*Diag: Chronic Pulmonary Disease
†Any Procedures Performed During Hospitalization
Num. Outpatient Visits
**Med: Supplement
Num. Days Since Last Inpatient Visit
**Med: Neuro Sedative
*Diag: Asthma
Insurance: Medicare/Medicaid
Insurance: Private
Num. Inpatient Visits in Study Period
Never Smoker
**Med: Gastrointestinal
**Med: Neuro Anticonvulsants
positive
negative
0
0.2
0.4
0.6
0.8
1.0
Standardized coef. magnitude

Figure 5.4 *(previous page)*: Important Predictors for 30-Day Readmissions in Sickle Cell Disease Selected by Logistic Regression Model. Normalized magnitude of a subset of the most important variables selected by the logistic regression model, averaged over the 100-independent train/test draws. The variables in blue are positively associated with the prediction outcome, and the variables in yellow are negatively associated with the prediction outcome. *Diag: Sickle Cell Genotype HbSS corresponds to the following ICD-9/ICD-10 codes: 282.62, 282.61, D57.0, D57.00, D57.01, D57.02; Diag: Chronic Pulmonary Disease corresponds to the following ICD-9/ICD-10 codes: 416.8, 416.9, 490.-505., 506.4, 508.1, 508.8, I27.8, I27.9, J40.-J47., J60.-J67., J68.4, J70.1, J70.3; Diag: Asthma corresponds to ICD-9 codes that start with 493. and the ICD-10 codes J44.0-J45. **Med: Cardiac Disease indicates whether a patient was prescribed with any cardiac medications during his or her stay, and this variable is used to indicate whether the patient has any cardiac comorbidities in addition to the ICD-9/ICD-10 coding. Similarly, Med: Overdose Reversal includes all medication that can be used to reverse a drug overdose; Med: Supplement includes all dietary supplements; Med: Neuro Sedative includes all anesthetics; Med: Gastrointestinal includes all drugs that can treat gastrointestinal diseases. ***ICD-10-PCS procedure code 30233N1 corresponds to "transfusion of nonautologous red blood cells into peripheral vein, percutaneous approach." ****Zip code 15221 corresponds to the borough of Wilkinsburg, PA, within the Pittsburgh metropolitan area. †Any procedures performed during the hospitalization is one of variables included by the LACE and HOSPITAL indices.

Model		Predicted Positive (%)	Predicted Negative (%)		
RF	True Positive (%)	26.29	32.72	Sensitivity (%)	44.5 ± 4.0
	True Negative (%)	6.05	34.94	Specificity (%)	85.2 ± 3.4
LR	True Positive (%)	24.85	32.77	Sensitivity (%)	43.1 ± 4.0
	True Negative (%)	6.26	36.12	Specificity (%)	85.2 ± 3.4
HOSPITAL	True Positive (%)	21.95	36.13	Sensitivity (%)	37.8 ± 3.9
	True Negative (%)	6.19	35.73	Specificity (%)	85.2 ± 3.4

Table 5.4: Out-of-sample Prediction Performance of the Random Forest and Logistic Regression Models Compared to HOSPITAL. Confusion matrices and corresponding sensitivities and specificities for the random forest and logistic regression classifier. A true positive (negative) case was determined as the admission did (not) result in a 30-day readmission and we correctly predicted so. The threshold of the HOSPITAL index is chosen to be 7 (Donzé et al. 2013), and the thresholds of RF and LR are chosen such that the specificities of these models match the specificity of the HOSPITAL index. Results are averaged over 100 independent train/test draws, where an average test set contains 134 patients and 1000 visits. Sensitivity and specificity are reported with 95% confidence intervals.

5.4 Discussion

This is the first study to apply ML algorithms to predict the hospital readmission rate in patients with SCD. Our model can be used at the point of discharge in a clinical setting. We have shown how the risk of 30-day readmission of a particular SCD patient can be estimated by preprocessing the EHR data associated with an inpatient admission using our data preprocessing steps, and then inputting the data into our pretrained model. Our models can be adapted to other regions and hospital systems by retraining the models to incorporate different zip codes. All variables included in our model are easily accessible through the EHR data.

The average age of SCD patients in our study cohort was 39.47 years. Since we excluded patients under 18 years old (given that we did not have access to our local pediatric EHR database), and the oldest patient in our cohort is above 90 years old compared to 56 years old in other studies (Brodsky et al. 2017), the average age in our study is slightly higher than those found in other studies (31.7 years old, Brodsky et al. 2017). We also found that the risk of rehospitalization is highest for the age group 18-29 in both Table 5.2 and Fig. 5.4, which is consistent with the results of a multi-state study of patients with SCD that revealed that acute care encounters and readmissions were most frequent in the 18-30 age group (Brousseau et al. 2010).

In our study, RF and LR appeared to be the best ML models in predicting hospital readmissions as seen in similar ML studies (Deschepper et al. 2019). To account for the fact that some patients might have a higher number of readmissions, we introduced a weighted RF model where each admission is weighted inversely by the total number of admissions incurred by the patient during the study period. The weighted RF model performs similar to the unweight RF model, indicating that the impact of those patients with frequent hospital admissions is small in our LR and RF models.

We discovered that ML methods were able to pick out additional variables specific to the SCD cohort that are underrepresented or absent in the traditional generalized hospital readmission scoring systems such as LACE (4 variables) and HOSPITAL (7 variables). All the variables from LACE and HOSPITAL were represented in our model, however, our models suggested the following variables were also predictive (Fig.s 5.3 and 5.4): labs (reticulocytes, platelets, bilirubin, white blood cells), demographic information (gender, zip code 15221), and SCD-specific comorbidities (chronic pulmonary disease, asthma).

For example, in our logistic regression model (Fig. 5.4), we observe that the majority of variables are in alignment with clinical experience and past studies. For instance, the number of inpatient visits over the past year, length of stay, and ED visits over the past 6 months are known to be risk factors for hospital readmissions (Brennan et al. 2015). The model found these variables positively correlated with higher risk of hospital readmissions. Conversely, having had a recent blood transfusion correlated negatively with the risk of hospital readmission in the model. These findings lend support to a previous study where the authors found that transfusion was associated with a reduced estimated odds ratio of inpatient mortality of 0.75 (95% CI: 0.57–0.99) and a decreased odds ratio of 30-day readmission of 0.78 (95% CI: 0.73–0.83) in the Truven Health MarketScan® Medicaid Databases (Nouraie and Gordeuk 2015).

Our random forest model (Fig. 5.3) contains a larger set of important features when compared with our logistic regression model (Fig. 5.4). In addition to the variables mentioned above, the random forest model also includes variables such as whether the patient has asthma or chronic obstructive pulmonary disease. However, in this model, the variables could contribute either positively or negatively to the readmission risk in our model. For example, it is possible that the age of the patient could contribute both positively and negatively towards the final readmission risk depending on the number of inpatient readmissions that the patient had in the past year. Thus, the features in Fig. 5.3 are not associated with any directions.

Our study underscores how ML may impact clinical care in SCD. However, since machine learning models test for correlations and not causations, further domain knowledge is needed to implement the model. Here we provide some examples of how such domain knowledge can be applied to exact meaningful interventions. For example, we found that zip code 15221, cardiac comorbidities (variable Med: Cardiac Disease), and age are significantly associated with hospital readmission risks among SCD patients. Since zip code 15221 is associated with a lower income community, and community resources may affect health outcomes, SCD clinics and comprehensive programs could mobilize resources to increase access to key healthcare resources for individuals with SCD residing in disadvantaged communities. For instance, SCD providers could establish strategic partnerships with community-based organizations and primary care providers in Federally Qualified Health Centers—community-based health care providers that receive funds from the Health Resources & Services Administration Health Center Program for primary care services in underserved areas—to provide behavioral health services, social services, and community outreach. In addition, health care plans and insurance providers may assist the SCD providers by assigning case managers and bolstering social work support for those patients with the highest readmission risk based on socioeconomic factors. Our ML model also identified medical factors for which both inpatient and outpatient interventions may be critical. We confirmed the emerging evidence that cardiac comorbidities significantly modulate the SCD phenotype (Gladwin and Sachdev 2012) by demonstrating their impact on 30-day readmission. Finally, age also emerged as an important factor in our model. This finding suggests that younger patients with SCD who may struggle navigating the challenging transition from pediatric to adult care could be engaged by partnering with the pediatric SCD providers to ensure continuity of care, ideally in a medical home setting. In summary, our study underscores the importance of identifying factors that affect 30-day readmission that can be targeted with a comprehensive, holistic, and medical home approach in SCD. This strategy is already bearing fruit for other chronic diseases that affect individuals throughout the lifespan (Jackson et al. 2013) and is likely to be critical for the vulnerable SCD community.

There are several limitations to our ML models. First, ICD coding may not always be reliable in EHR datasets (Futoma et al. 2015, Quan et al. 2005, Snyder et al. 2017). Since our dataset was de-identified, we were not able to verify if coding was correct by checking individual patients' EHR records. However, the majority of patients in our study cohort were diagnosed with SCD at least twice during the study period, increasing the likelihood that they were correctly identified as having SCD. To check the robustness of the SCD coding in our dataset, we re-performed two experiments with the following modifications: 1) with a subset of patients (identified in Table 5.2) with known sickle cell genotypes, and 2) with a subset of patients with at least two unplanned hospital readmissions. In both scenarios, we observed similar results. Figures 5.5 and 5.6 in § 5.7 illustrate the performance of our models as well as that of LACE and HOSPITAL indices under the above two scenarios. In addition, SCD genotypes were included as features in our models using ICD coding. In particular, our logistic regression model revealed that the genotype Hemoglobin SS (HbSS) was negatively associated with readmission risk (Fig. 5.4). There is evidence indicating that the coding of genotype HbSS is relatively accurate (with an error rate of 3%), but that the coding of genotype HbSC and HbS/B+ could be highly inaccurate (with error rates of 61% and 52% respectively), which is a limitation of coding in classifying genotype (Snyder et al. 2017). Thus, further research is needed to verify the impact of the latter two SCD genotypes on readmission risk. Second, socioeconomic factors and social determinants of health are inconsistently documented or not always accessible through the EHR alone (AlJuburi and Majeed 2013, Donzé et al. 2013, Kansagara et al. 2011, Cronin et al. 2019). Given this limitation, we relied on zip codes and insurance status as proxies of socioeconomic status (Table 5.5). Third, the data in our study might have contained missing admissions since patients might have been admitted into other hospitals outside the UPMC system. This limitation is similarly present in other studies (Xue et al. 2018, Shameer et al. 2017), and may be overcome by a more comprehensive data collection process (e.g., via survey), or by accessing multiple regional EHRs, to ensure the label of each visit is correct, but since our data was de-identified, we are unable to do so in this study and leave it to future studies. Finally, since SCD is a rare disease in the US according to NIH criteria, our sample size was relatively small. This precluded the use of more sophisticated ML models such as deep neural networks.

Our study demonstrates the feasibility of incorporating predictive analytical models with EHR data mining on a real-world data set to shed light on readmission patterns within a healthcare ecosystem; in particular, we showed the feasibility and potential of ML algorithms in predicting 30-day unplanned hospital readmissions for patients with SCD. Our best models, RF and LR, had relatively high predictive powers and could be useful in predicting 30-day readmissions within hospital systems. Thus, training ML models with disease specific variables can be valuable tools in predicting hospital readmission risk for SCD patients and may identify clinical variables not commonly included in readmission scores. If our model shows that a patient has a high readmission risk, then hospital resources can be allocated at point of discharge to include triaging with follow up visits and allocating specific resources to patient and family members to reduce readmission. In summary, we have developed a model that is more sensitive than existing models, suggesting that we can refine how we identify patients at high risk for readmission in SCD, but more investigation is needed to translate our findings into clinical interventions.

5.5 Preprocessing Tables

Table 5.5: Data Preprocessing. Description of the data preprocessing steps and the percentage of missing data. After preprocessing, we narrowed down the number of variables in our model to be 481. In the RF model, the vital sign variables are continuous, and we represent each vital sign variable using a tuple of size two with the first entry indicating whether the value of the variable is missing. This results in an overall vector representation of length 550. In the LR and SVM models, the vital sign variables were preprocessed into categorical variables and this results in an overall vector representation of length 565. In the third column, the number inside the parentheses is the size of the vector that we used to represent the corresponding features. The reasons that we used a larger vector to represent those features are due to 1) missing data 2) a categorical variable takes multiple values. In the fourth column, we described the percentage data missing overall. In the fifth column, we described the details on the variables included and the percentage of patients with this variable measured (if applicable) in the square brackets.

Variables (**C**ategorical /**R**eal valued)	Num. Vars. Pre	Num. Vars. Post (Rep.)	Missing Data	Variable Descriptions and Preprocessing Steps
Insurance providers (**C**)	50	4 (4)	None	Grouped insurance into 4 types: private, government, auto/employment, Medicare/Medicaid.
ICD-9/ICD-10 diagnosis codes (**C**)	3849	340 (340)	None	In addition to removing diagnosis codes that appeared less than 20 times, we also hand-picked 37 groups of diagnosis codes, including 3 sickle cell genotypes listed in Table 5.2 and 17 groups from the LACE index to calculate the Charlson comorbidity index score.
Procedures (**C**, **R**)	2808	25 (25)	454 (13.8%) visits had no procedures performed	Extracted whether any procedure was performed during the hospitalization (**C**) and the number of blood transfusions performed (**R**); removed the procedure codes (**C**) that appeared less than 20 times.

Lab tests (**C**)	2945	13 (78)	39 (1.2%) encounters had none of the 13 labs performed; see right for details on % encounters (out of 3299) have each of the 13 labs performed	Hand-picked 13 sickle cell related labs [% encounters have this test performed]: white blood cell count [98.5%], platelets count [98.5%], hemoglobin [15.5%], hematocrit [98.5%], reticulocytes count [77.1%], bilirubin [62.2%], lactic dehydrogenase (LDH) [58.3%] (tissue damage (i.e. anemia)), lactate blood [13.4%] (acid base imbalance i.e., lactic acidosis secondary to shock), creatinine [91.5%], bun/creatinine ratio [1.8%], creatinine clearance [0%], Pro BNP [0%], sodium (from the HOSPITAL index) [91.4%]. Each variable takes 6 categorical values and was represented by one-hot encoding. Table 5.6 describes the details of how those lab variables were extracted
Medication NDC codes (**C**)	4358	42 (43)	656 (19.8%) visits had no medication prescription; the rest had at least one prescription	Identified 553 unique drugs and grouped them into 42 categories based on the drug effect. An additional variable is added to represent whether any medication was prescribed during the inpatient admission.
Zip codes (**C**)	190	2 (2)	None	Removed the ones that appeared less than 20 times.
Smoking status (**C**)	10	5 (6)	685 (20.8%) encounters had no smoking status	Regrouped into: never smoker, former smoker, heavy tobacco smoker, light tobacco smoker, passive smoke exposure - never smoker
Hospital departments (**C**)	34	34 (34)	None	
Demographics (C, R)	2	2 (2)	None	[% patients have this demographic reported]: Gender (**C**) is binary [100%], age at encounter (**R**) [100%]

Vital signs (R, C)	7	7 (RF: 15; LR/ SVM: 30)	873 (26.5%) encounters had none of the vitals; see right for details	[% encounters have this vital ***taken***]: BMI (R or C:<18.5, 18.5-24.9, 25-29.9, 30-34.9, 35-39.9, ≥40) [71.5%], BP_systolic (R or C: <90, 90-120, >120) [60.0%], BP_diastolic (R or C: <60, 60-80, >80) [59.7%], pulse (R or C: <60, 60-100, >100) [59.7%], temperature (R or C: ≤35C/95F, (35C, 38C)/(95F, 100.4F), ≥38C/100.4F) [59.3%], respiratory_rate (R or C: <12, 12-18, >18) [59.6%], BP_position (**C**) [1.3%]
Other variables included (**R**)	7	7 (7)	NA	Length of stay, number of outpatient visits, number of ED visits, number of ED visits in the past 6 months, the number of days since the last inpatient visit, number of inpatient visits in study period, number of inpatient visits in the past year.

Table 5.6: Lab variables included in the study. The percentages of Reticulocytes result that were normal, low, and high in this study was 37.6%, 0.9%, and 61.5%, respectively.

Lab Category	Included variable names	Excluded variable names
White blood cell	WBC, WHITE BLOOD CELLS, WBC COUNT, WBC & OTHER NUCLEATED CELLS	WHITE BLOOD CELLS-URINE >5 WBC/HPF (POC), WBC Esterase, Rare WBCs present no organisms present, No WBCs or organisms present, No WBCs present few gram positive cocci in pairs, WBC - fluid, WBC Morphology, WBC clumps, Fecal WBC, IMMATURE WBC FORMS
Platelets	PLATELETS, PLATELET COUNT	PLATELET MORPHOLOGY, HEPARIN PF4 PLATELET ANTIBODY, HEPARIN PLATELET AB, GIANT PLATELETS, PLATELET ESTIMATE, LARGE PLATELETS, PLATELET FUNCTION P2Y12, PLATELET SUFFICIENCY, MEAN PLATELET VOLUME, RAPID PRA(PLATELETS), PLATELET FUNCTION INTERP.

Hemoglobin	HEMOGLOBIN F, RAPID HEMOGLOBIN S, HEMOGLOBIN C. CRYSTALS, HEMOGLOBIN S, HEMOGLOBIN C, HEMOGLOBIN A2, HEMOGLOBIN-PLASMA, TOTAL HEMOGLOBIN, THB (HEMOGLOBIN)	METHEMOGLOBIN &&, % OXYHEMOGLOBIN, HEMOGLOBIN (POCT), HEMOGLOBIN - MIXED VENOUS, METHEMOGLOBIN - MIXED VENOUS, % REDUCED HEMOGLOBIN, ATYPICAL HEMOGLOBIN, HEMOCUE HEMOGLOBIN (POCT), METHEMOGLOBIN - VENOUS, GLYCOSYLATED HEMOGLOBIN, METHEMOGLOBIN, CARBOXYHEMOGLOBIN, “Hemoglobin, qual”, HEMOGLOBIN A, HEMOGLOBIN A1, HEMOGLOBIN A1C, HEMOGLOBIN CAPILLARY (POC), BEDSIDE HEMOGLOBIN POCT, HEMOGLOBIN-ARTERIAL, HEMOGLOBIN-VENOUS, CALC. HEMOGLOBIN ISTAT
Hematocrit	HEMATOCRIT, HEMATOCRIT(HCT)	HEMATOCRIT DERIVED, HEMATOCRIT DERIVED - MIXED VEN, HEMATOCRIT(HCT) MANUAL PCV &&, HEMATOCRIT (POCT), HEMATOCRIT-BODY FLUID (HCT), HEMATOCRIT ISTAT
Reticulocytes	ABSOLUTE RETICULOCYTES, RETICULOCYTES, RETICULOCYTES-MANUAL	METHOD IMMATURE RETICULOCYTE FRACTION
Bilirubin	TOTAL BILIRUBIN, DIRECT BILIRUBIN, BILIRUBIN UNCONJUGATED	BILIRUBIN-URINE, BILIRUBIN - URINE (POC), BILIRUBIN CONFIRMATION, BILIRUBIN UNCONJUGATED, OTHER TOTAL BILIRUBIN
Lactic dehydrogenase (LDH) (tissue damage (i.e., anemia))	LACTIC DEHYDROGENASE, LACTIC DEHYDROGENASE(LDH), OTHER LACTIC DEHYDROGENASE(LD)	“LDH, ASCITES FLUID”

Lactate blood (acid base imbalance i.e., lactic acidosis secondary to shock)	LACTATE, LACTATE BLOOD, LACTATE WHOLE BLOOD	LACTATE CSF, LACTATE ISTAT
Creatinine	CREATININE, CREATININE, WHOLE BLOOD, RANDOM URINE CREATININE, "CREATININE, RANDOM URINE"	"CREATININE, JP DRAINAGE", CREATININE VENOUS ISTAT, FLUID CREATININE, CREATININE POCT, URINE PROTEIN/CREATININE RATIO, PROTEIN/CREATININE RATIO, URINE CREATININE, TOTAL CREATININE 24 HR URINE, CREATININE ISTAT
Bun/creatinine ratio	BUN/CREATININE RATIO	ALBUMIN/CREATININE RATIO
Creatinine clearance	CREATININE CLEARANCE, CREATININECLEARANCE(ADULT)	CREATININE CLEAR.(CHILDREN'S)
BNP	"PRO BNP, N-TERMINAL"	
Sodium	SODIUM(NA), SODIUM(NA) WHOLE BLOOD, SODIUM NA WHOLE BLOOD, SODIUM ARTERIAL BLOOD GAS	STOOL SODIUM (AKA NASTOOL), SODIUM ISTAT, URINE SODIUM(NA), TOTAL SODIUM(NA) 24HR URINE, SODIUM (NA) (POCT)

5.6 Descriptions of the Four ML Algorithms, LACE Index, and HOSPITAL Index

Logistic Regression and Support Vector Machine Classifications. The logistic regression and support vector machine are two well-known machine learning algorithms for classification, both with linear decision boundaries. Logistic regression learns a logic function that maps the input features (our predictor candidates) to the target label (whether this admission results in a readmission). For our own algorithm, to award logit function with sparse weights, we added L1 penalty to our logistic regression algorithm, where the inverse of regularization strength was set to 0.05. Support vector machine learns a hyperplane that "best" separates data (input features) with the opposite labels. Since in practice the input data is not always linearly separable (as in our case), we first transform our data using the Laplace RBF kernel, commonly used under no prior knowledge on the data, and set the regularization parameter to 1. Both methods were implemented using the scikit-learn package.

Random Forest Classification. The random forest model is a well-known machine learning algorithm for classification. A random forest is made up of multiple decision trees that each make simple classification decisions based on relatively few variables. These trees are created (or "trained") with different, randomly drawn subsets of variables so that it is likely that no two trees are identical. Given a new sample, each tree is traversed top-down until a set of training samples is reached at the bottom. Using the forest as a whole for classification amounts to having the multiple decision trees "vote" on a label (in this case, case or control), where each tree's vote is made from the labels of the bottom set of training samples. Thus, in a binary classification problem, given an input sample, the likelihood outputting 1 is calculated by taking the average of the probability of outputting 1 among all decision trees. The probability of outputting 1 in a single decision tree is calculated by dividing the number of samples of the class 1 by the total number of samples in a leaf. We use this likelihood, along with the true label corresponding to the input sample, to compute the C-statistics of our model.

For our own algorithm, each random forest consisted of 400 decision trees, each with a maximum depth of 15. The model was trained using the scikit-learn package. Importance scores were also calculated using the same package. On an individual "tree" of the random forest, the importance score of any variable used in constructing the tree is defined as the proportion of the training set that lies in the 'leaves' of nodes utilizing that variable (variables not used in constructing the tree are assigned a score of zero); then the overall importance score for a variable is the average of its importance scores on each tree.

The Weighted Random Forest Classification. Unlike the random forest model described above, where each sample contributes equally to training the model, in the weighted random forest model, each sample is assigned with a predefined weight. To make sure that each patient is given even consideration in the model, let x_i denote the total number of unplanned inpatient admissions that patient i had during the study period. Then, the weight of each inpatient admission of patient i is is calculated by $1/x_i$. Similar to the random forest model above, each random forest in this model consisted of 400 decision trees, each with a maximum depth of 15. The model was also trained using the scikit-learn package.

The LACE and HOSPITAL Indices. The LACE index includes 4 variables, namely the length of stay, acuity of the admission (a binary variable where 1 indicates the admission was through the emergency room, 0 indicates an elective admission for planned intervention), Charlson comorbidity index score, and the number of emergency department visits within the six months before admission (van Walraven et al. 2010, Donzé et al. 2013), and returns a 30-day readmission risk score. In particular, the Charlson comorbidity index score was calculated using the ICD code at discharge (See Table 5.5). Since elective admissions were removed from our sample, every admission in our study cohort has an acuity of 1. The HOSPITAL index includes 7 variables, namely the length of stay, the number of hospital admissions in the twelve months before admission, the admission type, whether any procedure is performed during the stay, patient blood sodium level prior to discharge, whether the patient is discharged from an oncology service, and patient blood hemoglobin level prior to discharge, and returns a 30-day readmission risk score.14 Similar to LACE, the patient admission type is non-elective for all patients; furthermore, because of the logistics pertaining to inpatient SCD care in the UPMC system, all patients were discharged from an internal medicine service. The AUCs of these two indices were calculated the same way as those of ML methods.

5.7 Additional Experimental Results with different patient inclusion criteria

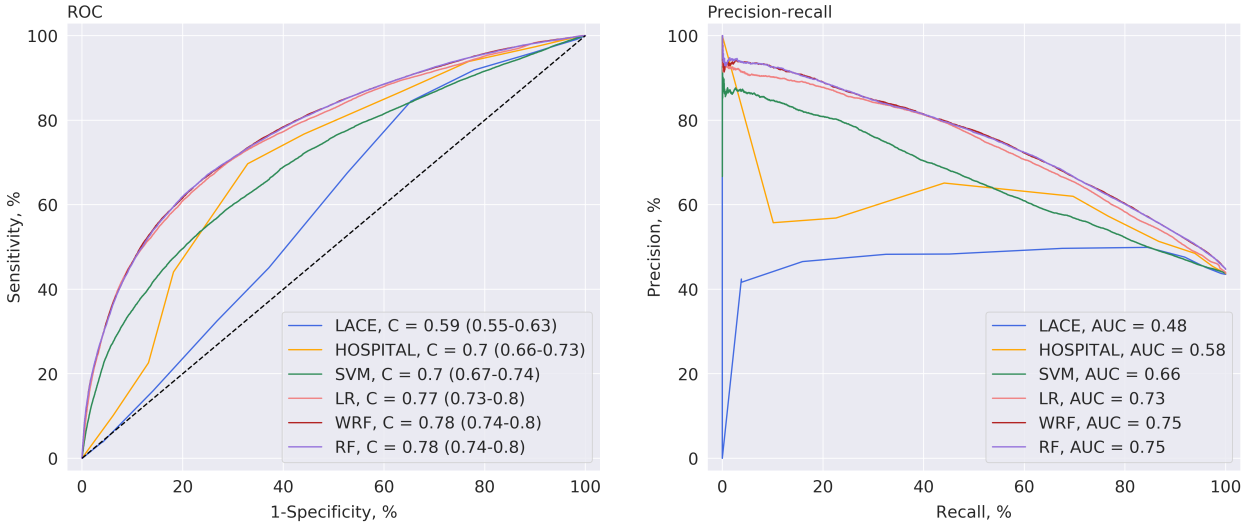


Figure 5.5: Performance Metrics of Machine Learning Models for Predicting 30-Day Readmissions in Patients with Known Sickle Cell Genotypes. Two performance metrics measured out-of-sample and averaged over 100 independent train/test draws. (A) Receiver operating characteristic curves, and corresponding area under the curve; also known as the C-statistic. (B) Precision-recall curves. Patients with unknown sickle cell genotypes were removed from the study cohort described in the main sections. This results in 314 patients, and 153 of these patients had 30-day readmissions. The total number of inpatient admissions made by these 314 patients is 2914.

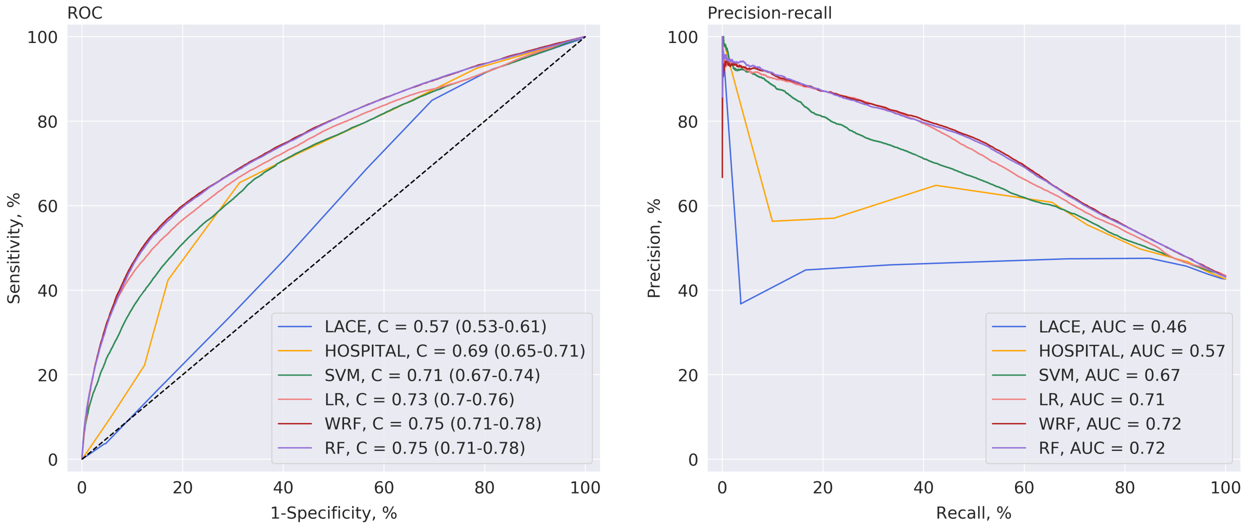


Figure 5.6: Performance Metrics of Machine Learning Models for Predicting 30-Day Readmissions in Sickle Patients with Two or More Admissions. Two performance metrics measured out-of-sample and averaged over 100 independent train/test draws. (A) Receiver operating characteristic curves, and corresponding area under the curve; also known as the C-statistic. (B) Precision-recall curves. Patients with only one inpatient admission during the study period were removed from the study cohort described in the main sections. This results in 286 patients, and 195 of these patients had 30-day readmissions. The total number of inpatient admissions made by these 284 patients is 3167.

Bibliography

Adler M, Heeringa B (2008) Approximating optimal binary decision trees. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, 1–9 (Springer).

Adzika VA, Glozah FN, Ayim-Aboagye D, Ahorlu CS (2017) Socio-demographic characteristics and psychosocial consequences of sickle cell disease: the case of patients in a public hospital in ghana. *Journal of Health, Population and Nutrition* 36(1):4.

Alaa AM, van der Schaar M (2017) Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 3424–3432.

Alagoz O, Chhatwal J, Burnside ES (2013) Optimal policies for reducing unnecessary follow-up mammography exams in breast cancer diagnosis. *Decision Analysis* 10(3):200–224.

AlJuburi G, Majeed A (2013) Trends in hospital admissions for sickle cell disease in england. *Journal of Public Health* 35(1):179–179.

Amalakuhan B, Kiljanek L, Parvathaneni A, Hester M, Cheriyath P, Fischman D (2012) A prediction model for copd readmissions: catching up, catching our breath, and improving a national problem. *Journal of Community Hospital Internal Medicine Perspectives* 2(1):9915.

Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434):444–455.

Arkin EM, Meijer H, Mitchell JS, Rappaport D, Skiena SS (1998) Decision trees for geometric models. *International Journal of Computational Geometry & Applications* 8(03):343–363.

Armitage P (1950) Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society. Series B* 12(1):137–144.

ASAM (2016) Opioid addiction treatment. URL https://www.asam.org/docs/default-source/publications/asam-opioid-patient-piece_-5bopt2-5d_3d.pdf.

Athey S, Chetty R, Imbens G, Kang H (2016) Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index.

Awasthi P, Balcan MF, Long PM (2017) The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)* 63(6):1–27.

Ayer T, Alagoz O, Stout NK (2012) Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research* 60(5):1019–1034.

Ayer T, Alagoz O, Stout NK, Burnside ES (2015) Heterogeneity in women's adherence and its role in optimal breast cancer screening policies. *Management Science* 62(5):1339–1362.

Ayvaci MU, Alagoz O, Burnside ES (2012) The effect of budgetary restrictions on breast cancer diagnostic decisions. *Manufacturing & Service Operations Management* 14(4):600–617.

Azar Y, Gamzu I (2011) Ranking with submodular valuations. *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, 1070–1079 (SIAM).

Azar Y, Gamzu I, Yin X (2009) Multiple intents re-ranking. *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 669–678 (ACM).

Balcan M, Beygelzimer A, Langford J (2006) Agnostic active learning. *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, 65–72.

Ballas SK, Lusardi M (2005) Hospital readmission for adult acute sickle cell painful episodes: frequency, etiology, and prognostic significance. *American journal of hematology* 79(1):17–25.

Banerjee S, Karri SPK, Chatterjee S, Pal M, Paul RR, Chatterjee J (2016) Multimodal diagnostic segregation of oral leukoplakia and cancer. *Systems in Medicine and Biology (ICSMB), 2016 International Conference on*, 67–70 (IEEE).

Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–973.

Bareinboim E, Pearl J (2013) A general algorithm for deciding transportability of experimental results. *Journal of causal Inference* 1(1):107–134.

Barnett PG (2009) Comparison of costs and utilization among buprenorphine and methadone patients. *Addiction* 104(6):982–992.

Baser O, Chalk M, Fiellin DA, Gastfriend DR (2011) Cost and utilization outcomes of opioid-dependence treatments. *The American Journal of Managed Care* 17:S235–48.

Bates S, Sesia M, Sabatti C, Candes E (2020) Causal inference in genetic trio studies. *arXiv preprint arXiv:2002.09644* .

Benenson I, Jadotte Y, Echevarria M (2017) Factors influencing utilization of hospital services by adult sickle cell disease patients: a systematic review. *JBI database of systematic reviews and implementation reports* 15(3):765–808.

Best MG, Sol N, Kooi I, Tannous J, Westerman BA, Rustenburg F, Schellen P, Verschueren H, Post E, Koster J, et al. (2015) Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer cell* 28(5):666–676.

Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM, et al. (2014) Detection of circulating tumor dna in early-and late-stage human malignancies. *Science translational medicine* 6(224):224ra24–224ra24.

Breiman L (2001) Random forests. *Machine learning* 45(1):5–32.

Brennan JJ, Chan TC, Killeen JP, Castillo EM (2015) Inpatient readmissions and emergency department visits within 30 days of a hospital admission. *Western Journal of Emergency Medicine* 16(7):1025.

Brodsky MA, Rodeghier M, Sanger M, Byrd J, McClain B, Covert B, Roberts DO, Wilkerson K, DeBaun MR, Kassim AA (2017) Risk factors for 30-day readmission in adults with sickle cell disease. *The American journal of medicine* 130(5):601–e9.

Brom H, Carthon JMB, Ikeaba U, Chittams J (2020) Leveraging electronic health records and machine learning to tailor nursing care for patients at high risk for readmissions. *Journal of Nursing Care Quality* 35(1):27–33.

Brousseau DC, Owens PL, Mosso AL, Panepinto JA, Steiner CA (2010) Acute care utilization and rehospitalizations for sickle cell disease. *Jama* 303(13):1288–1294.

Brown SE, Weisberg DF, Balf-Soran G, Sledge WH (2015) Sickle cell disease patients with and without extremely high hospital use: pain, opioids, and coping. *Journal of Pain and Symptom Management* 49(3):539–547.

Bubeck S, Munos R, Stoltz G (2009) Pure exploration in multi-armed bandits problems. *International conference on Algorithmic learning theory*, 23–37 (Springer).

Campbell CM, Edwards RR (2012) Ethnic differences in pain and pain management. *Pain Management* 2(3):219–230.

cancerorg (2021) American cancer society: Cancer facts & statistics. URL https://cancerstatisticscenter.cancer.org/.

Cassandra A, Littman ML, Zhang NL (1997) Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. *Uncertainty in Artificial Intelligence (UAI)*, 54–61.

Castro RM, Nowak RD (2007) Minimax bounds for active learning. *International Conference on Computational Learning Theory*, 5–19 (Springer).

Cevik M, Ayer T, Alagoz O, Sprague BL (2018) Analysis of mammography screening policies under resource constraints. *Production and Operations Management* 27(5):949–972.

Chakaravarthy VT, Pandit V, Roy S, Sabharwal Y (2009) Approximating decision trees with multiway branches. *International Colloquium on Automata, Languages, and Programming*, 210–221 (Springer).

Chalana H, Kundal T, Gupta V, Malhari AS (2016) Predictors of relapse after inpatient opioid detoxification during 1-year follow-up. *Journal of Addiction* 2016:1–7.

Chan KA, Woo JK, King A, Zee BC, Lam WJ, Chan SL, Chu SW, Mak C, Tse IO, Leung SY, et al. (2017) Analysis of plasma epstein–barr virus dna to screen for nasopharyngeal cancer. *New England Journal of Medicine* 377(6):513–522.

Chawla S, Gergatsouli E, Teng Y, Tzamos C, Zhang R (2019) Learning optimal search algorithms from data. *arXiv preprint arXiv:1911.01632* .

Chen G, Kim S, Taylor JM, Wang Z, Lee O, Ramnath N, Reddy RM, Lin J, Chang AC, Orringer MB, et al. (2011) Development and validation of a quantitative real-time polymerase chain reaction classifier for lung cancer prognosis. *Journal of Thoracic Oncology* 6(9):1481–1487.

Chen M, Hao Y, Hwang K, Wang L, Wang L (2017) Disease prediction by machine learning over big data from healthcare communities. *Ieee Access* 5:8869–8879.

Chen Q, Ayer T, Chhatwal J (2018) Optimal m-switch surveillance policies for liver cancer in a hepatitis c–infected population. *Operations Research* 66(3):673–696.

Chen Y, Hassani SH, Karbasi A, Krause A (2015) Sequential information maximization: When is greedy near-optimal? *Conference on Learning Theory*, 338–363.

Chen Y, White RS, Tangel V, Noori SA, Gaber-Baylis LK, Mehta ND, Pryor KO (2019) Sickle cell disease and readmissions rates after lower extremity arthroplasty: a multistate analysis 2007–2014. *Journal of comparative effectiveness research* 8(6):403–422.

Cheol Jeong I, Bychkov D, Searson PC (2018) Wearable devices for precision medicine and health state monitoring. *IEEE Transactions on Biomedical Engineering* 66(5):1242–1258.

Chernoff H (1959) Sequential design of experiments. *The Annals of Mathematical Statistics* 30(3):755–770.

Chirikov VV, Shaya FT, Onukwugha E, Mullins CD, dosReis S, Howell CD (2017) Tree-based claims algorithm for measuring pretreatment quality of care in medicare disabled hepatitis c patients. *Medical care* 55(12):e104–e112.

Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A, et al. (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359(6378):926–930.

Cole SR, Hernán MA (2008) Constructing inverse probability weights for marginal structural models. *American journal of epidemiology* 168(6):656–664.

Connock M, Juarez-Garcia A, Jowett S, Frew E, Liu Z, Taylor R, Fry-Smith A, Day E, Lintzeris N, Roberts T, et al. (2007) Methadone and buprenorphine for the management of opioid dependence: a systematic review and economic evaluation. *Health Technology Assess* 11(9):1–171.

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 22(1):173–203.

Cortes C, Vapnik V (1995) Support-vector networks. *Machine learning* 20(3):273–297.

Cosmic (2019) Cosmic - catalogue of somatic mutations in cancer. URL https://cancer.sanger.ac.uk/.

Crist RC, Clarke TK, Ang A, Ambrose-Lanci LM, Lohoff FW, Saxon AJ, Ling W, Hillhouse MP, Bruce RD, Woody G, et al. (2013) An intronic variant in oprd1 predicts treatment outcome for opioid dependence in african-americans. *Neuropsychopharmacology* 38(10):2003–2010.

Cronin RM, Hankins JS, Byrd J, Pernell BM, Kassim A, Adams-Graves P, Thompson A, Kalinyak K, DeBaun M, Treadwell M (2019) Risk factors for hospitalizations and readmissions among individuals with sickle cell disease: results of a us survey study. *Hematology* 24(1):189–198.

Cuzick J, Thorat MA, Andriole G, Brawley OW, Brown PH, Culig Z, Eeles RA, Ford LG, Hamdy FC, Holmberg L, et al. (2014) Prevention and early detection of prostate cancer. *The lancet oncology* 15(11):e484–e492.

Dasgupta S (2005) Analysis of a greedy active learning strategy. *Advances in neural information processing systems*, 337–344.

Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84(1):151–161.

Deschepper M, Eeckloo K, Vogelaers D, Waegeman W (2019) A hospital wide predictive model for unplanned readmission using hierarchical icd data. *Computer methods and programs in biomedicine* 173:177–183.

Donzé J, Aujesky D, Williams D, Schnipper JL (2013) Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA internal medicine* 173(8):632–638.

Du D, Hwang FK, Hwang F (2000) *Combinatorial group testing and its applications*, volume 12 (World Scientific).

Eastwood B, Strang J, Marsden J (2017) Effectiveness of treatment for opioid use disorder: a national, five-year, prospective, observational study in england. *Drug & Alcohol Dependence* 176:139–147.

Eckert C, Nieves-Robbins N, Spieker E, Louwers T, Hazel D, Marquardt J, Solveson K, Zahid A, Ahmad M, Barnhill R, et al. (2019) Development and prospective validation of a machine learning-based risk of readmission model in a large military hospital. *Applied clinical informatics* 10(2):316.

Epstein DH, Willner-Reid J, Vahabzadeh M, Mezghanni M, Lin JL, Preston KL (2009) Real-time electronic diary reports of cue exposure and mood in the hours before cocaine and heroin craving and use. *Archives of General Psychiatry* 66(1):88–94.

Erenay FS, Alagoz O, Said A (2014) Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing & Service Operations Management* 16(3):381–400.

Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, Radich J, Anderson G, Hartwell L (2003) The case for early detection. *Nature Reviews Cancer* 3(4):243–252.

Even-Dar E, Mannor S, Mansour Y (2002) Pac bounds for multi-armed bandit and markov decision processes. *International Conference on Computational Learning Theory*, 255–270 (Springer).

Fatseas M, Denis C, Massida Z, Verger M, Franques-Rénéric P, Auriacombe M (2011) Cue-induced reactivity, cortisol response and substance use outcome in treated heroin dependent individuals. *Biological Psychiatry* 70(8):720–727.

Fatseas M, Serre F, Alexandre JM, Debrabant R, Auriacombe M, Swendsen J (2015) Craving and substance use among patients with alcohol, tobacco, cannabis or heroin addiction: A comparison of substance-and person-specific cues. *Addiction* 110(6):1035–1042.

Ferguson MK, Wang J, Hoffman PC, Haraf DJ, Olak J, Masters GA, Vokes EE (2000) Sex-associated differences in survival of patients undergoing resection for lung cancer. *The Annals of thoracic surgery* 69(1):245–249.

Fogarty CB, Small DS (2016) Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *Journal of the American Statistical Association* 111(516):1820–1830.

Frei-Jones MJ, Field JJ, DeBaun MR (2009) Risk factors for hospital readmission within 30 days: a new quality measure for children with sickle cell disease. *Pediatric blood & cancer* 52(4):481–485.

Futoma J, Morris J, Lucas J (2015) A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics* 56:229–238.

Gan K, Li AA, Lipton ZC, Tayur S (2020) Causal inference with selectively deconfounded data. *arXiv preprint arXiv:2002.11096* .

Gladwin MT, Sachdev V (2012) Cardiovascular abnormalities in sickle cell disease. *Journal of the American College of Cardiology* 59(13):1123–1133.

Golovin D, Krause A (2011) Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *J. Artif. Intell. Res.* 42:427–486.

Gorodeski EZ, Ishwaran H, Kogalur UB, Blackstone EH, Hsich E, Zhang Zm, Vitolins MZ, Manson JE, Curb JD, Martin LW, et al. (2011) Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the women's health initiative. *Circulation: Cardiovascular Quality and Outcomes* 4(5):521–532.

Grella CE, Lovinger K (2011) 30-year trajectories of heroin and other drug use among men and women sampled from methadone treatment in california. *Drug & Alcohol Dependence* 118(2):251–258.

Guillory A, Bilmes JA (2011) Simultaneous learning and covering with adversarial noise. *ICML*.

Gupta V, Han BR, Kim SH, Paek H (2020) Maximizing intervention effectiveness. *Management Science* .

Hannah L, Blei D, Powell W (2010) Dirichlet process mixtures of generalized linear models. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 313–320 (JMLR Workshop and Conference Proceedings).

Hanneke S, Yang L (2015) Minimax analysis of active learning. *The Journal of Machine Learning Research* 16(1):3487–3602.

Hanneke S, et al. (2014) Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning* 7(2-3):131–309.

Hartman E, Grieve R, Ramsahai R, Sekhon JS (2015) From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(3):757–778.

Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4):1161–1189.

Holland PW (1986) Statistics and causal inference. *Journal of the American statistical Association* 81(396):945–960.

Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS (2011) Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes* 4(1):39–45.

Huang Y, Valtorta M (2006) Pearl's calculus of intervention is complete. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, 217–224.

Humphreys K, Malenka RC, Knutson B, MacCoun RJ (2017) Brains, environments, and policy responses to addiction. *Science* 356(6344):1237–1238.

Im S, Nagarajan V, Zwaan RVD (2016) Minimum latency submodular cover. *ACM Transactions on Algorithms (TALG)* 13(1):13.

Isom JD, Meyn SP, Braatz RD (2008) Piecewise linear dynamic programming for constrained pomdps. *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 1, 291–296.

Jackson GL, Powers BJ, Chatterjee R, Prvu Bettger J, Kemper AR, Hasselblad V, Dolor RJ, Irvine RJ, Heidenfelder BL, Kendrick AS, et al. (2013) The patient-centered medical home: a systematic review. *Annals of internal medicine* 158(3):169–178.

Jemal A, Siegel R, Xu J, Ward E (2010) Cancer statistics, 2010. *CA: a cancer journal for clinicians* 60(5):277–300.

Jerant AF, Johnson JT, Sheridan C, Caffrey TJ, et al. (2000) Early detection and treatment of skin cancer. *American family physician* 62(2):357–386.

Jia S, Navidi F, Nagarajan V, Ravi R (2019) Optimal decision tree with noisy outcomes. *Advances in neural information processing systems* .

Joynt KE, Jha AK, et al. (2012) Thirty-day readmissions—truth and consequences. *N Engl j med* 366(15):1366–1369.

Kallus N, Puli AM, Shalit U (2018) Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 10888–10897.

Kamath G, Tzamos C (2019) Anaconda: A non-adaptive conditional sampling algorithm for distribution testing. *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 679–693.

Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S (2011) Risk prediction models for hospital readmission: a systematic review. *JAMA* 306(15):1688–1698.

Kelty E, Hulse G (2017) Fatal and non-fatal opioid overdose in opioid dependent patients treated with methadone, buprenorphine or implant naltrexone. *International Journal of Drug Policy* 46:54 – 60.

Kim D, Lee J, Kim KE, Poupart P (2011) Point-based value iteration for constrained pomdps. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1968–1974.

Kim Y, Jeon J, Mejia S, Yao CQ, Ignatchenko V, Nyalwidhe JO, Gramolini AO, Lance RS, Troyer DA, Drake RR, et al. (2016) Targeted proteomics identifies liquid-biopsy signatures for extracapsular prostate cancer. *Nature communications* 7.

Kleber HD, Weiss RD, Anton Jr RF, George TP, Greenfield SF, Kosten TR, O'Brien CP, Rounsaville BJ, Strain EC, Ziedonis DM, Hennessy G, Connery HS (2006) Practice guideline for the treatment of patients with substance use disorders 111–124, URL http://www.psych.org/.

Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M (2002) *Logistic regression* (Springer).

Knudson AG (2001) Two genetic hits (more or less) to cancer. *Nature Reviews Cancer* 1(2):157–162.

Kolodny A, Frieden TR (2017) Ten steps the federal government should take now to reverse the opioid addiction epidemic. *JAMA* 318(16):1537–1538.

Kosaraju SR, Przytycka TM, Borgstrom R (1999) On an optimal split tree problem. *Workshop on Algorithms and Data Structures*, 157–168 (Springer).

Krause A, McMahan HB, Guestrin C, Gupta A (2008) Robust submodular observation selection. *Journal of Machine Learning Research* 9(Dec):2761–2801.

Krebs E, Min JE, Evans E, Li L, Liu L, Huang D, Urada D, Kerr T, Hser YI, Nosyk B (2017) Estimating state transitions for opioid use disorders. *Medical Decision Making* 37(5):483–497.

Kroll RR, McKenzie ED, Boyd JG, Sheth P, Howes D, Wood M, Maslove DM (2017) Use of wearable devices for post-discharge monitoring of icu patients: a feasibility study. *Journal of Intensive Care* 5(1):64.

Kuroki M, Pearl J (2014) Measurement bias and effect restoration in causal inference. *Biometrika* 101(2):423–437.

Laurent H, Rivest RL (1976) Constructing optimal binary decision trees is np-complete. *Information processing letters* 5(1):15–17.

Lautieri A (2019) How long do opiates stay in your system? hydrocodone, morphine, heroin. URL https://americanaddictioncenters.org/prescription-drugs/how-long-in-system.

Lee J, Kim GH, Poupart P, Kim KE (2018) Monte-carlo tree search for constrained pomdps. *Advances in Neural Information Processing Systems (NeurIPS)*, 7934–7943.

Linder C (2019) To stop opioid overdoses, startups are developing fitbit-like wearables. lots of them. URL https://www.post-gazette.com/business/tech-news/2019/02/05/This-startup-is-working-on-a-wearable-device-to-prevent-opioid-overdose/stories/201901250007.

Lipton ZC, Kale DC, Wetzel R (2016) Modeling missing data in clinical time series with rnns. *arXiv preprint arXiv:1606.04130* .

Liu M, Oxnard G, Klein E, Swanton C, Seiden M, Liu MC, Oxnard GR, Klein EA, Smith D, Richards D, et al. (2020) Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free dna. *Annals of Oncology* 31(6):745–759.

Lo-Ciganic WH, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, Donohue JM, Cochran G, Gordon AJ, Malone DC, et al. (2019) Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA Network Open* 2(3):e190968–e190968.

Lorden G (1977) Nearly-optimal sequential tests for finitely many parameter values. *The Annals of Statistics* 1–21.

Louizos C, Shalit U, Mooij JM, Sontag D, Zemel R, Welling M (2017) Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, 6446–6456.

Luo SX, Levin FR (2017) Towards precision addiction treatment: New findings in co-morbid substance use and attention-deficit hyperactivity disorders. *Current Psychiatry Reports* 19(3):14.

Machado RF, Barst RJ, Yovetich NA, Hassell KL, Kato GJ, Gordeuk VR, Gibbs JSR, Little JA, Schraufnagel DE, Krishnamurti L, et al. (2011) Hospitalization for pain in patients with sickle cell disease treated with sildenafil for elevated trv and low exercise capacity. *Blood, The Journal of the American Society of Hematology* 118(4):855–864.

Maitra P, Caughey M, Robinson L, Desai PC, Jones S, Nouraie M, Gladwin MT, Hinderliter A, Cai J, Ataga KI (2017) Risk factors for mortality in adult patients with sickle cell disease: a meta-analysis of studies in north america and europe. *Haematologica* 102(4):626–636.

Mannor S, Tsitsiklis JN (2004) The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research* 5(Jun):623–648.

Manterola L, Guruceaga E, Pérez-Larraya JG, González-Huarriz M, Jauregui P, Tejada S, Diez-Valle R, Segura V, Samprón N, Barrena C, et al. (2014) A small noncoding rna signature found in exosomes of gbm patient serum as a diagnostic tool. *Neuro-oncology* 16(4):520–527.

Marlin BM, Kale DC, Khemani RG, Wetzel RC (2012) Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, 389–398.

Marsden J, Eastwood B, Ali R, Burkinshaw P, Chohan G, Copello A, Burn D, Kelleher M, Mitcheson L, Taylor S, et al. (2014) Development of the addiction dimensions for assessment and personalised treatment (adapt). *Drug & Alcohol Dependence* 139:121–131.

Marshall DA, Burgos-Liz L, IJzerman MJ, Osgood ND, Padula WV, Higashi MK, Wong PK, Pasupathy KS, Crown W (2015) Applying dynamic simulation modeling methods in health care delivery research—the simulate checklist: report of the ispor simulation modeling emerging good practices task force. *Value in Health* 18(1):5–16.

Mavrotas G (2009) Effective implementation of the ε-constraint method in multi-objective mathematical programming problems. *Applied mathematics and computation* 213(2):455–465.

McCaffrey DF, Ridgeway G, Morral AR (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* 9(4):403.

McLellan AT, Alterman AI, Metzger DS, Grissom GR, Woody GE, Luborsky L, O'brien CP (1994) Similarity of outcome predictors across opiate, cocaine, and alcohol treatments: role of treatment services. *Journal of Consulting and Clinical Psychology* 62(6):1141–1158.

McLellan AT, Kushner H, Metzger D, Peters R, Smith I, Grissom G, Pettinati H, Argeriou M (1992) The fifth edition of the addiction severity index. *Journal of Substance Abuse Treatment* 9(3):199–213.

Mehari A, Gladwin MT, Tian X, Machado RF, Kato GJ (2012) Mortality in adults with sickle cell disease and pulmonary hypertension. *Jama* 307(12):1254–1256.

Miao W, Geng Z, Tchetgen Tchetgen EJ (2018) Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* 105(4):987–993.

Miller A, Hoogstraten B, Staquet M, Winkler A (1981) Reporting results of cancer treatment. *Cancer* 47(1):207–214.

Miratrix LW, Wager S, Zubizarreta JR (2018) Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika* 105(1):103–114.

Mitzenmacher M, Upfal E (2017) *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis* (Cambridge university press).

Morral AR, Iguchi MY, Belding MA, Lamb RJ (1997) Natural classes of treatment response. *Journal of Consulting and Clinical Psychology* 65(4):673.

Morse S, Bride BE (2017) Decrease in healthcare utilization and costs for opioid users following residential integrated treatment for co-occurring disorders. *Healthcare* 5(3):54.

Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, Negahban SN, Krumholz HM (2016) Analysis of machine learning techniques for heart failure readmissions. *Circulation: Cardiovascular Quality and Outcomes* 9(6):629–640.

Naghshvar M, Javidi T (2013) Active sequential hypothesis testing. *The Annals of Statistics* 41(6):2703–2738.

Navidi F, Kambadur P, Nagarajan V (2020) Adaptive submodular ranking and routing. *Operations Research* .

Neyman J (1923) Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10:1–51.

NIDA (2020) Fiscal year 2020 budget information—congressional justification for national institute on drug abuse. https://www.drugabuse.gov/about-nida/legislative-activities/budget-information.

Nikolaev AG, Jacobson SH, Cho WKT, Sauppe JJ, Sewell EC (2013) Balance optimization subset selection (boss): An alternative approach for causal inference with observational data. *Operations Research* 61(2):398–412.

Nosyk B, MacNab YC, Sun H, Fischer B, Marsh DC, Schechter MT, Anis AH (2009) Proportional hazards frailty models for recurrent methadone maintenance treatment. *American Journal of Epidemiology* 170(6):783–792, URL http://dx.doi.org/10.1093/aje/kwp186.

Nouraie M, Gordeuk VR (2015) Blood transfusion and 30-day readmission rate in adult patients hospitalized with sickle cell disease crisis. *Transfusion* 55(10):2331–2338.

Nowak RD (2009) Noisy generalized binary search. *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, 1366–1374.

O'Rourke N, Edwards R (2000) Lung cancer treatment waiting times and tumour growth. *Clinical Oncology* 12(3):141–144.

O'Toole TP, Pollini RA, Ford D, Bigelow G (2006) Physical health as a motivator for substance abuse treatment among medically ill adults: is it enough to keep them in treatment? *Journal of Substance Abuse Treatment* 31(2):143–150.

Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, et al. (2004) Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304(5676):1497–1500.

Pearl J (1995) Causal diagrams for empirical research. *Biometrika* 82(4):669–688.

Pearl J (2000) *Causality: models, reasoning and inference*, volume 29 (Springer).

Pearl J (2010) Brief report: On the consistency rule in causal inference:" axiom, definition, assumption, or theorem?". *Epidemiology* 872–875.

Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, Anagnostou V, Fiksel J, Cristiano S, Papp E, et al. (2017) Direct detection of early-stage cancers using circulating tumor dna. *Science translational medicine* 9(403):eaan2415.

Poupart P, Malhotra A, Pei P, Kim KE, Goh B, Bowling M (2015) Approximate linear programming for constrained partially observable markov decision processes. *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 29.

Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA (2005) Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care* 1130–1139.

Raıssouli M, Jebril IH (2010) Various proofs for the decrease monotonicity of the schatten's power norm, various families of r n- norms and some open problems. *Int. J. Open Problems Compt. Math* 3(2):164–174.

Rayner E, van Gool IC, Palles C, Kearsey SE, Bosse T, Tomlinson I, Church DN (2016) A panoply of errors: polymerase proofreading domain mutations in cancer. *Nature Reviews Cancer* .

Razavi P, Li BT, Abida W, Aravanis A, Jung B, Shen R, Hou C, De Bruijn I, Gnerre S, Lim RS, et al. (2017) Performance of a high-intensity 508-gene circulating-tumor dna (ctdna) assay in patients with metastatic breast, lung, and prostate cancer. *J Clin Oncol* 35(18 Suppl).

Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling* 7(9-12):1393–1512.

Robins JM (2000) Robust estimation in sequentially ignorable missing data and causal inference models. *Journal of the American Statistical Association on Bayesian Statistical Science* 6–10.

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89(427):846–866.

Rosenbaum PR (1987) Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* 74(1):13–26.

Rosenbaum PR (2002) Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association* 97(457):183–192.

Rosenbaum PR (2011) A new u-statistic with superior design sensitivity in matched observational studies. *Biometrics* 67(3):1017–1027.

Rosenbaum PR (2014) Weighted m-statistics with superior design sensitivity in matched observational studies with multiple controls. *Journal of the American Statistical Association* 109(507):1145–1158.

Rosenbaum PR, Rosenbaum P, Briskman (2010) *Design of observational studies*, volume 10 (Springer).

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rosenman E, Owen AB, Baiocchi M, Banack H (2018) Propensity score methods for merging observational and experimental datasets. *arXiv preprint arXiv:1804.07863* .

Rotnitzky A, Robins JM, Scharfstein DO (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association* 93(444):1321–1339.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688.

Sargan JD (1958) The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society* 393–415.

Sayre SL, Schmitz JM, Stotts AL, Averill PM, Rhoades HM, Grabowski JJ (2002) Determining predictors of attrition in an outpatient substance abuse program. *The American Journal of Drug and Alcohol Abuse* 28(1):55–72.

Schackman BR, Leff JA, Polsky D, Moore BA, Fiellin DA (2012) Cost-effectiveness of long-term outpatient buprenorphine-naloxone treatment for opioid dependence in primary care. *Journal of General Internal Medicine* 27(6):669–676.

Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94(448):1096–1120.

Serre F, Fatseas M, Debrabant R, Alexandre JM, Auriacombe M, Swendsen J (2012) Ecological momentary assessment in alcohol, tobacco, cannabis and opiate dependence: a comparison of feasibility and validity. *Drug & Alcohol Dependence* 126(1):118–123.

Serre F, Fatseas M, Denis C, Swendsen J, Auriacombe M (2018) Predictors of craving and substance use among patients with alcohol, tobacco, cannabis or opiate addictions: commonalities and specificities across substances. *Addictive Behaviors* 83:123–129.

Serre F, Fatseas M, Swendsen J, Auriacombe M (2015) Ecological momentary assessment in the investigation of craving and substance use in daily life: a systematic review. *Drug & Alcohol Dependence* 148:1–20.

Settles B (2009) Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Shameer K, Johnson KW, Yahi A, Miotto R, Li L, Ricks D, Jebakaran J, Kovatch P, Sengupta PP, Gelijns S, et al. (2017) Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using mount sinai heart failure cohort. *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, 276–287 (World Scientific).

Shen C, Li X, Li L, Were MC (2011) Sensitivity analysis for causal inference using inverse probability weighting. *Biometrical Journal* 53(5):822–837.

Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, LaVange L, Marinac-Dabic D, Marks PW, Robb MA, et al. (2016) Real-world evidence—what is it and what can it tell us. *N Engl J Med* 375(23):2293–2297.

Shi X, Miao W, Nelson JC, Tchetgen EJT (2018) Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding.

Shpitser I, Pearl J (2006a) Identification of conditional interventional distributions. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, 437–444.

Shpitser I, Pearl J (2006b) Identification of joint interventional distributions in recursive semi-markovian causal models. *Proceedings of the National Conference on Artificial Intelligence*, volume 21, 1219 (Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999).

Siegel RL, Miller KD, Jemal A (2015) Cancer statistics, 2015. *CA: a cancer journal for clinicians* 65(1):5–29.

Sinha R (2008) Chronic stress, drug use, and vulnerability to addiction. *Annals of the new York Academy of Sciences* 1141(1):105–130.

Skolnick P (2018) The opioid epidemic: crisis and solutions. *Annual Review of Pharmacology and Toxicology* 58:143–159.

Smallwood RD, Sondik EJ (1973) The optimal control of partially observable markov processes over a finite horizon. *Operations Research* 21(5):1071–1088.

Snyder AB, Lane PA, Zhou M, Paulukonis ST, Hulihan MM (2017) The accuracy of hospital icd-9-cm codes for determining sickle cell disease genotype. *Journal of rare diseases research & treatment* 2(4):39.

Sondik EJ (1971) The optimal control of partially observable markov processes. Technical report, Stanford Univ Calif Stanford Electronics Labs.

Stuart EA, Cole SR, Bradshaw CP, Leaf PJ (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2):369–386.

Suen Sc, Brandeau ML, Goldhaber-Fiebert JD (2017) Optimal timing of drug sensitivity testing for patients on first-line tuberculosis treatment. *Health Care Management Science* 21(4):632–646.

Suzuki M, Mitoma H, Yoneyama M (2017) Quantitative analysis of motor status in parkinson's disease using wearable devices: From methodological considerations to problems in clinical applications. *Parkinson's Disease* 2017:1–9.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. (2019) Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research* 47(D1):D941–D947.

Termorshuizen F, Krol A, Prins M, van Ameijden EJC (2005) Long-term outcome of chronic drug use: the amsterdam cohort study among drug users. *American Journal of Epidemiology* 161(3):271–279.

Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, Rashidi P, Pardalos P, Momcilovic P, Bihorac A (2016) Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one* 11(5):e0155705.

Tian J, Pearl J (2002) A general identification condition for causal effects. *Aaai/iaai*, 567–573.

Undurti A, How JP (2010) An online algorithm for constrained pomdps. *IEEE International Conference on Robotics and Automation*, 3966–3973.

Valant V, Lindemer E, Ghafari N (2018) Hey,charlie home page. URL https://heycharlie.org/.

Van der Laan MJ, Rose S (2011) *Targeted learning: causal inference for observational and experimental data* (Springer Science & Business Media).

van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, Austin PC, Forster AJ (2010) Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj* 182(6):551–557.

VanderWeele TJ, Ding P (2017) Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine* 167(4):268–274.

Varatharajan R, Manogaran G, Priyan M, Sundarasekar R (2019) Wearable sensor devices for early detection of alzheimer disease using dynamic time warping algorithm. *Cluster Computing* 21(1):681–690.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wald A (1945) Sequential tests of statistical hypotheses. *The annals of mathematical statistics* 16(2):117–186.

Walraven E, Spaan MT (2019) Point-based value iteration for finite-horizon pomdps. *Journal of Artificial Intelligence Research* 65:307–341.

Wang Y, Singh A (2016) Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Weinreich M, Nguyen OK, Wang D, Mayo H, Mortensen EM, Halm EA, Makam AN (2016) Predicting the risk of readmission in pneumonia. a systematic review of model performance. *Annals of the American Thoracic Society* 13(9):1607–1614.

White CC (1991) A survey of solution techniques for the partially observed markov decision process. *Annals of Operations Research* 32(1):215–230.

Williamson A, Darke S, Ross J, Teesson M (2006) The effect of persistence of cocaine use on 12-month outcomes for the treatment of heroin dependence. *Drug & Alcohol Dependence* 81(3):293–300.

Wilson-Frederick SM, Hulihan M, Anderson KK (2019) Prevalence of sickle cell disease among medicaid beneficiaries in 2012. *CMS Office of Minority Health Data Highlight* .

Wines JD, Saitz R, Horton NJ, Lloyd-Travaglini C, Samet JH (2007) Overdose after detoxification: a prospective study. *Drug & Alcohol Dependence* 89(2):161–169.

Xue Y, Liang H, Norbury J, Gillis R, Killingworth B (2018) Predicting the risk of acute care readmissions among rehabilitation inpatients: A machine learning approach. *Journal of biomedical informatics* 86:143–148.

Zarkin GA, Dunlap LJ, Hicks KA, Mamo D (2005) Benefits and costs of methadone treatment: results from a lifetime simulation model. *Health Economics* 14(11):1133–1150.

Zhang J, Denton BT, Balasubramanian H, Shah ND, Inman BA (2012) Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management* 14(4):529–547.

Zhang NL, Liu W (1996) Planning in stochastic domains: Problem characteristics and approximation. Technical report, Technical Report HKUST-CS96-31, Hong Kong University of Science and Technology.

Zhao Q, Small DS, Bhattacharya BB (2017) Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *arXiv preprint arXiv:1711.11286* .