Planning for HIV Screening, Testing and Care at the Veterans Health Administration

Sarang Deo
Indian School of Business, Gachibowli, Hyderabad, India 500032
(sarang_deo@isb.edu)
Kumar Rajaram, Sandeep Rath, Uday Karmarkar
UCLA Anderson School of Management, Los Angeles CA 90095
(kumar.rajaram@anderson.ucla.edu, sandeep.rath.2015@anderson.ucla.edu
uday.karmarkar@anderson.ucla.edu)
Matthew Goetz
Veteran's Health Administration, Greater Los Angeles Station
(matthew.goetz@va.gov)

Abstract

We analyzed the planning problem for HIV screening, testing and care. This problem consists of determining the optimal fraction of patients to be screened in every period as well as the optimumstaffing level at each part of the health care system to maximize the total health benefits to the patients measured by Quality-Adjusted Life-Years (QALYs) gained. We modeled this problem as a nonlinear mixed integer programming program comprising of disease progression (the transition of the patients across health states), system dynamics (the flow of patients in different health states across various parts of the health care delivery system), budgetary and capacity constraints. We applied the model to the Greater Los Angeles (GLA) station in the Veterans Health Administration (VHA)system. We found that a Center for Disease Control recommended routine screening policy in which all patients visiting the system are screened for HIV irrespective of risk factors may not be feasible due to budgetary constraints. Consequently, we used the model to develop and evaluate managerially relevant policies within existent capacity and budgetary constraints to improve upon the current risk based screening policy of screening only high risk patients. Our computational analysis showed that the GLA station can achieve substantial increase (20% to300%) in the QALYs gained by using these policies over risk based screening. The GLA station has already adapted two of these policies that could yield better patient health outcomes over the next few years. In addition, our model insights have influenced the decision making process at this station.

1. Introduction

Veterans Health Administration (VHA), one of the components of the Veterans Administration, is the largest integrated healthcare provider in the United States of America (USA). The VHA is funded by the federal government and serves the medical and social support needs of over 8 million active duty and honorably discharged veterans over their entire lifetime. The VHA provides these services through 128 stations. For the purpose of this paper, we shall focus on the Greater Los Angeles (GLA) station, as the unit of analysis because of our close working relationship with its key decision makers.

Electronic copy available at: http://ssrn.com/abstract=2331107

The VHA is the largest provider of HIV care in the USA. As of 2011, the VHA reported over 25,271 HIV infected patients, an increase of 3.7% from 2007. The VHA is also a leader in quality of care provided to HIV infected patients with high adherence to the Department of Health and Human Services clinical guidelines across all regions.An important aspect of HIV care is early diagnosis and treatment which is known to lower cost and improvepatient outcomes (Palella et al. 2003). In addition, this reduces the incidence of secondary complications which are very costly to treat if HIV itself is not treated in a timely manner (Schackmanet al. 2006). Prior studies at the VHA (Nayak et al. 2012) show that a major factorimpeding the early diagnosis and treatment of HIV is the policy of*risk-based*screening. Under this policy, patients are tested for HIV only if they display certain risk factors such as injection drug use, or if they present symptoms of opportunistic infections. Owens et al. (2007) found that only 36% of at risk patients had ever been tested for HIV. The main operational barriers cited for insufficient coverage of screening and late diagnosis of HIV infection were constraints on provider time and insufficient capacity of trained counselors (Goetz et al. 2008a).

An alternative policyrecommended by the Centers for Disease Control (CDC)is to implement *routine* HIV screening, in which a patient visiting the health care facility would be offered an HIV test irrespective of risk factors or symptoms. Several recent studies in the public health literature have found that such routine HIV screening is "cost-effective"[1]compared to risk based testing even in settings with very low prevalence of HIV (Paltiel et al. 2005). In 2009, the VHA proposed to implement the routine screening policy across its stations[2]. Consequently, the management at the GLA station wanted to understand if such a policy would be feasible given their capacity and budgetary constraints, and if necessary,was willing to consider alternate policies to improve upon their current risk based screening policy.In response, we developed an optimization model to achieve these goalsat the GLA station. Consistent with the mission of the VHA of providing high quality care over the lifetime of veterans, the objective of this model is to maximize the total QALYs of all the patients at this station. To achieve this objective, this model determines the optimal fraction of patients to be screened (i.e., offered the test) and also determines the optimum staffing levels at different parts or locations of the station. This model explicitly captures patient flow and the associated disease progression through system dynamics constraints. In addition, it also incorporates budget and capacity constraints.

[1]A policy or interventionis said to be "cost effective" if the Quality Adjusted Life Years (QALYs) gained due to that intervention costs less than $109,000 to $297,000 per QALY gained. (http://www.cdc.gov/hiv/prevention/ongoing/costeffectiveness/). The term QALYs is commonly used in the health economics and health policy literature to assess the value of a medical intervention in terms of the number of years at a particular quality level added due to the intervention(Dolan et al. 2005).

[2]http://www.va.gov/vhapublications/ViewPublication.asp?pub_ID=2056

Electronic copy available at: http://ssrn.com/abstract=2331107

We first used this model to evaluate the current risk based screening policy and the proposed routine screening policy at the GLA station. We found that the cost-effective routine screening policy was not feasible in the current budgetary environment at this station. Therefore, we developed four other policies within the framework of our model that improved upon the current risk based screening policy. An extensive computational analysis provided a benchmark value for each policy and provided guidance in terms of the fraction of patients to be screened in every period as well as the number of health care workers that need to be staffed at each part of the system in order to implement a policy. Thus, unlike conventional cost effective analysis, our approach provided a feasible plan that can be implemented.

Optimization based models have been used to evaluate prevention and treatment policies for HIV at differentdecision makinglevels(Kahn et al. 1998; Rauner et al. 2001). Population level studies evaluate the cost effectiveness of policy interventions (Zaric et al. 2000; Long et al. 2010), while studies at an individual patient level optimize clinical decision making to maximize patient welfare (Shechter et al. 2008; Roberts et al. 2010). Health care systems facethe problem of integrating cost effective policies with clinical decisions subject to organizational and budgetary constraints. Blount et al. (1997), Zaric and Brandeau (2001) andBrandeau et al. (2003) evaluate general formulations of this problem with budget constraints to decide optimal intervention for prevention of infectious diseases. Their approximations lead to formulations that can be solved by linear programming and convex optimization techniques. More recently, Kucuzyaciki et al. (2011) and Deo et al. (2013) combine clinical models of disease progression for chronic diseases with operational models of the health system. However, none of these papers consider different partsof the health care system with capacity constraints and do not jointly optimize screening and staffing decisions, which are the key features of the decision problem faced by the VHA.

Our paper makes the following contributions. First, it models a very relevant but complex problem at the interface of operations management and public health. It then develops methods for the efficient computation of bounds and managerially relevant solutions for this problem. Second, to the best of our knowledge, this is the first planning model which determines the fraction of patients that need to be screened along with the staffing requirements at screening, testing and care, while including disease progression and flow of patients in different health states across various parts of a constrained health care system. Third, we explicitly consider capacity and budget constraints and illustrate their impact on screening and staff allocation decisions. Fourth, we apply the model to data collected from the GLA station to analyze various policies. Our computational analysis shows that GLA station can achieve substantial increase (20% to 300%) in the QALYs gained by using these policies and our model provides guidance for its effective implementation.Fifth, the insights from our model have influenced planning decisions at this station. In addition, two policies have been used at the GLA station and our analysis provides the basis to extend and enhance these policies.

The remainder of the paper is organized as follows. In Section 2, we describe the health care system, patient health states, disease progression and system dynamics.These form the basis of our optimization model, which is formulated in Section 3. We also discuss structural properties, construct an upper bound and develop four policies that serve as lower bounds for this model. In Section 4, we describe various primary and secondary sources of data used in the model. Section 5 analyzes several policies for HIV screening, testing and care that can be evaluated within the framework of our model. Section 6 describes the application and qualitative impact of this work.

2. Problem Description

The GLA station is one of the largest and the most complex stations in the VHA consisting of 3 ambulatory care centers, a tertiary care facility and 10 community based clinics. The GLA serves veterans residing in Los Angeles, Kern, Santa Barbara, Ventura and San Louis Obispo counties. The GLA station management recommendedthat we conduct a station level analysis because it was difficult to estimate the budget for individual facilities within the station. Further, the management felt that such an analysis could lead to effective staff reallocationbecause there was considerable flexibility in adjusting the staffing levels across facilities within a station.From a managerial perspective, these aspects were considered more important than any potential downside due to loss of granularity in terms of patient flow and staffing.

As discussed before, the primary benefit of routine screening is early diagnosis of HIV positive patients and their connection to care before they become symptomatic. This benefit arises from the fact that the health care cost of asymptomatic HIV patients (including HIV treatment and other hospitalization) is much lower and their quality of life is much better than that of symptomatic HIV patients (Kaplan et al. 2009). In order to capture this effect, we constructed a compartmental model of patients with each compartment corresponding to a combination of the health state of the patientsand part of the health care system to which they belong. Below, we describe the health care system, patient health states, disease progression, and system dynamics.

2.1 Health Care System

Based on our discussions with the stationmanagement, we divided the health care system at the station into three distinct parts: 1)*primary care* (facilities such as outpatient clinics and hospitals where patients are screened or are offered an HIV test and blood samples are collected if they agree to be tested), 2) *laboratory*(a central location where samples collected during screening are tested), and 3) *infectious disease specialty care* (where HIV positive patients are referred for monitoring or treatment). Primary and specialty care could be staffed by up to three worker types: physicians, nurses and counselors, while the

laboratory is only staffed by the laboratory technician.Staffing levels are fixed during the budget horizon of one yearto provide certaintyand foster a stable work environment for all their staff.

To provide a precise definition of the health care system, let $\tau \in [T] = \{1, 2, \ldots, T\}$denote the budget periods each corresponding to a yearand let $t \in \mathcal{M}_\tau = \{1 + 12(\tau - 1), \ldots, 12\tau\}$ index the set of discrete time periods corresponding to a month within the budget period. Further, let $k \in \mathcal{W} = \{phys, nurse, couns, lab\}$ index the set of worker types and $l \in \mathcal{L} = \{P, L, S\}$ index the set of parts or locations where P denotes primary care facility, L denotes laboratory and S denotes infectious diseases specialty care. Each location l is staffed by $n_{k,l}$ health care workers of type k, each of whom earns a wage $w_k$in each period and spends a total of $y_{k,l}$ time units on average with the patient. Since the health care workers have other tasks associated with other diseases and conditions, we assume that the total time available with the resource of type$k$in location $l$for the HIV routine screening program is limited and denoted by$A_{k,l}$.

2.2 Patient Health States

Following earlier work in the modeling of disease progression in HIV patients (Freedberg et al. 1998; Mauskopf et al.2005),we use different ranges of CD4 cell count[3], and the presence or absence of Opportunistic Infections (OI) to define a set of health states of HIV infected patients. In addition, we include uninfected and dead as two additional health states.Table 1 below provides the definition of the resulting 14 health states based on CD4 count range and their associated states of OI. These states are indexed by $i$and$j$in the model.

[Insert Table 1 here]

In addition, the VHA identifies incoming patients as either high-risk or low-risk depending on their observable characteristics such as previous Hepatitis B or C infection, injection drug use or homelessness. These risk categories are indexed by $r \in \mathcal{R} = \{1, 2\}$, where$r = 1$, signifies patients of higher risk of infection of HIV and $r = 2$, signifies those with a lower risk of infection. At the GLA station, 25% of the patients were classified as high risk while the remaining 75% were classified as low risk (Goetz et al. 2013).

2.3 Disease Progression

In single patient models, the transition between health states is typically modeled as a discrete time Markov chain in which the probability of transitioning from state i to state j is conditionally independent of the history of earlier transitions. However, this approach is analytically intractable for a multi-

[3]$CD4^+$T helper cells are white blood cells essential to the human immune system and are usually expressed as number of cells per milliliter. Patients infected with HIV show reduced number of CD4 cells and a lower number of CD4 indicates a greater progression of the infection.

periodaggregate or population-level model like ours, which also considers multiple parts of the health care system while optimizing screening and staff allocation decisions. Hence, we approximate the disease progression model by using deterministic transition rates in which we assume that a fixed fraction of the number of patients move from one health state to the other in each period[4].This deterministic approximation of transition rates is reasonable here since the unit of our analysis is the GLA station and the population of patients in each state is relatively large. We use$\theta_{r,\omega}^{i,j}$to denote the fraction of patients in health state $i$that move to health state j in one month. This fraction depends on the patient risk category r, and the treatment status $\omega \in \mathcal{S} = \{treat, untreat\}$, where $treat$ refers to undergoing antiretroviral treatment and$untreat$ represents not undergoing treatment respectively.

Four processes govern the transition across health states: 1) HIV infection, 2) HIV infection progression (treated and untreated), 3) Opportunistic infection(OI), and 4) OI recovery.We used clinical data to estimate the transition rates associated with each of these processes separately. For certain transitions that require more than one process simultaneously, we assumed that the rate of one process does not depend on the other. Details on the calculations of the transition rates are provided in the Electronic Companion.

2.4 System Dynamics

In this section, we describe the system dynamics obtainedby combining disease progression with patient flows to represent how patients move across different health states as well as various parts of the health care system over time. In particular, we track the number of patients in each risk category r, each health state i at each location l in each time period$t$.Figure 1 shows the flow of patients through various parts of the health care system.

[Insert Fig 1 here]

Primary Care–Screening

The process starts with patients who are unaware of their HIV status, whom we call unscreened patients. Let $U_{r,t}^{i}$denote the total number of unscreened patients in risk category r, health state i and at time period t.All patients with an opportunistic infection($i \in \mathcal{I}_o = \{7, 8, \ldots, 13\}$) are immediately offered the HIV testand their acceptance rate is 100%. A fraction α of the remaining asymptomatic patients who do not have OI ($i \in \mathcal{I}_w = \{0, 1, \ldots, 6\}$) visit a primary care facility in period $t$for other conditions.Let $S_{r,t}$represent the fraction of patients of risk category r in period t that are screened or offered the HIV test. Afraction β of these patients accept the test. The number of unscreenedpatientsin the next time period, $U_{r,t+1}^{i}$is given by:

[4] A similar approach is used in mathematical epidemiology to model the spread of infectious diseases in the population (Anderson et. al. 2002).

$$U_{r,t+1}^{i} = \left(\sum_{j \in \mathcal{J}_w} \theta_{r,untreat}^{j,i} \left(1 - \alpha\beta S_{r,t}\right) U_{r,t}^{j} \right) + N_{r,t+1}^{i} + R_{r,t}^{0} \theta_{r,untreat}^{0,i} \qquad \forall r, i, t \qquad (1)$$

The first term $\left(\sum_{j \in \mathcal{J}_w} \theta_{r,untreat}^{j,i} \left(1 - \alpha\beta S_{r,t}\right) U_{r,t}^{j}\right)$of this equation is derived by summing three types of patient flows shown in Figure 1: a) the asymptomatic patients who do not visit the clinic, b) those who visit and do not get screened, c) those who visit, get selected for a test and refuse to be tested. This sum is appropriately weighted by the rates of transition from state j to state i as determined by the disease progression model. The second term, $N_{r,t+1}^{i}$ is the number of new patients in health state i and risk category r who enter in period$(t+1)$ as shown in Figure 1. The third term ($R_{r,t}^{0} \theta_{r,untreat}^{0,i}$) is the number of uninfected patients who receive a negative HIV test at the beginning of period t and join the pool of unscreened population in the next period.

Laboratory - Testing

The blood samples collected from patients who accept the offered testare then sent to the lab where the actual test is conducted and the results are communicated back to the patient. Here, we allow for a lag between the collection of the sample and return of the results due to congestion at the lab. Let $W_{r,t+1}^{i}$represent the number of patients in health state i, risk category$r$who are waiting to receive their results at the beginning of the period$t+1$in the laboratory. This is given by:

$$W_{r,t+1}^{i} = \sum_{j \in \mathcal{J}} W_{r,t}^{j} \theta_{r,untreat}^{j,i} + \sum_{j \in \mathcal{J}_w} \alpha\beta S_{r,t}\, U_{r,t}^{j} \theta_{r,untreat}^{j,i} + \sum_{j \in \mathcal{J}_o} U_{r,t}^{j} \theta_{r,untreat}^{j,i} - \sum_{j \in \mathcal{J}} R_{r,t}^{j} \theta_{r,untreat}^{j,i} \qquad \forall r, i, t \qquad (2)$$

$W_{r,t+1}^{i}$ consists of four terms. The first term ($\sum_{j \in \mathcal{J}} W_{r,t}^{j} \theta_{r,untreat}^{j,i}$) represents the number of patients waiting at the beginning of period t who have undergone disease progression, where $\mathcal{J} = \mathcal{J}_o \cup \mathcal{J}_w$. The second term, ($\sum_{j \in \mathcal{J}_w} \alpha\beta S_{r,t}\, U_{r,t}^{j} \theta_{r,untreat}^{j,i}$) represents the number of asymptomatic patients who accept the test offer at the beginning of period t. The third term, ($\sum_{j \in \mathcal{J}_o} U_{r,t}^{j} \theta_{r,untreat}^{j,i}$), represents the number of symptomatic patients who directly proceed to testing. The fourth term ($\sum_{j \in \mathcal{J}} R_{r,t}^{j} \theta_{r,untreat}^{j,i}$) represents the patients who receive their results and who either exit the system as their tests are negative (i.e., $j = 0$) or who are now transferred to care (i.e.,$j \neq 0$). As before, multiplication by $\theta_{r,untreat}^{j,i}$in each term represents disease progression in one period.

Specialty Care – Monitoring and Treatment

Patients who receive positive test results are connected to infectious diseases specialty care for monitoring and treatment. Again, we allow for a lag between the receipt of results and being connected to care. Let $I_{r,t}^{i}$ denote the number of patients of risk category r and health state $i$who are initiated in care. Of these, depending on the stage of their disease progression, $IM_{r,t}^{i}$ are initiated under monitoring while $ID_{r,t}^{i}$ are immediately initiated on treatment. Let $E_{r,t+1}^{i}$ denote the number of patients at the beginning of the period $t+1$ who are waiting to be enrolled in care. This is given by:

$$E_{r,t+1}^{i} = \sum_{j \in \mathcal{J}/\{0\}} R_{r,t}^{j} \theta_{r,untreat}^{j,i} + \sum_{j \in \mathcal{J}/\{0\}} E_{r,t}^{j} \theta_{r,untreat}^{j,i} - \sum_{j \in \mathcal{J}/\{0\}} IM_{r,t}^{j} \theta_{r,untreat}^{j,i} - \sum_{j \in \mathcal{J}/\{0\}} ID_{r,t}^{j} \theta_{r,treat}^{j,i} \qquad \forall r,i,t \qquad (3)$$

The first term, ($\sum_{j \in \mathcal{J}/\{0\}} R_{r,t}^{j} \theta_{r,untreat}^{j,i}$) is the number of patients who received positive HIV test results at the beginning of period *t*. The second term, ($\sum_{j \in \mathcal{J}/\{0\}} E_{r,t}^{j} \theta_{r,untreat}^{j,i}$), is the number of patients who were waiting to be enrolled into care at the beginning of period t. The third and fourth term ($\sum_{j \in \mathcal{J}/\{0\}} IM_{r,t}^{j} \theta_{r,untreat}^{j,i}$, $\sum_{j \in \mathcal{J}/\{0\}} ID_{r,t}^{j} \theta_{r,treat}^{j,i}$) are the number of people who were enrolled at period t into monitoring and treatment respectively. Patients who are enrolled into treatment now undergo disease progression under the parameter $\theta_{r,treat}^{j,i}$ instead of $\theta_{r,untreat}^{j,i}$.

The decision to initiate patients under monitoring or under treatment depends on the health state of the patient and current clinical guidelines described in Section 4.2. We use a binary indicator parameter z^{i} to capture the clinical decision whether all patients at health state i are initiated under treatment ($z^{i}=1$) or monitoring ($z^{i}=0$). Then, the number of patients who are initiated into treatment and monitoring at time period *t* are given by the following equations:

$$ID_{r,t}^{i} = I_{r,t}^{i} z^{i} \qquad \forall r,i,t \qquad (4)$$

$$IM_{r,t}^{i} = I_{r,t}^{i}\left(1 - z^{i}\right) \qquad \forall r,i,t \qquad (5)$$

Next, consider$M_{r,t+1}^{i}$, the number of patients of risk category r under monitoring in state i at the beginning of period $t+1$. This is given by:

$$M_{r,t+1}^{i} = \sum_{j \in \mathcal{J}/\{0\}} M_{r,t}^{j} \theta_{r,untreat}^{j,i} - \sum_{j \in \mathcal{J}/\{0\}} M_{r,t}^{j} z^{j} \theta_{r,treat}^{j,i} + \sum_{j \in \mathcal{J}/\{0\}} IM_{r,t}^{j} \theta_{r,untreat}^{j,i} \qquad \forall r,i,t \qquad (6)$$

The first term in Equation (6) represents the number of patients in health state i who remain under monitoring at the beginning of period t, the second term represents who enter treatment from monitoring and the third term represents the newly diagnosed patients who enter care under monitoring.

Finally, let $D^{i}_{r,t+1}$represent the number of patients under treatment in state i at the beginning of period t. This is given by:

$$D^{i}_{r,t+1} = \sum_{j \in \mathcal{J}/\{0\}} D^{j}_{r,t}\, \theta^{j,i}_{r,treat} + \sum_{j \in \mathcal{J}/\{0\}} M^{j}_{r,t} z^{j}\, \theta^{j,i}_{r,treat} + \sum_{j \in \mathcal{J}/\{0\}} ID^{j}_{r,t}\, \theta^{j,i}_{r,treat} \qquad \forall r, i, t \qquad (7)$$

The first term in equation (7) represents the number of patients under treatment in period t in a particular health state, the second term denotes the number of patients who enter treatment from the pool of monitored patients and the third term is the number of newly diagnosed patients who enter treatment.

In formulating the system dynamics (1) through (7), we have made the following simplifying assumptions. First, once patients enter the system and are tested, they can exit the system only if they are uninfected or if they die. Second, all primary care locations fully comply with the screening policy. Third, the treatment protocol is well defined and is followed by all physicians at the infectious diseases specialty. These assumptions were validated by prior internal studies at the GLA station.Given the health care system, patient health states, disease progression and system dynamics the overall objective of the GLA station is to maximize the aggregated Quality Adjusted Life Years (QALYs) across all patients in the system. This can be done by appropriately choosing the screening fraction and consequently the number of patients to be screened, tested and cared in every period and by determining the staffing level at each part of the health care system to execute this choice. While doing this the station faces organizational constraints relating to capacity and budget availability. We next develop an optimization model for this decision problem.

3. Model

In this section, we start by describing the objective function and the organizational constraints related to budget and capacity. These together with the previously described system dynamics form a discrete time planning model. We characterize key properties of this model, and use them to develop an upper bound which can be employed to evaluate the quality of any given solution. Finally, we develop managerially relevant heuristics or policies to solve this model. Table 2 summarizes all notations that are used in the model including those that have already been introduced in the previous section.

[Insert Table 2 here]

3.1 Objective Function

In accordance with the existing literature on economic evaluation of health interventions and programs (Dolan et al.2005) and discussions with senior administrators at the station, we choose the objective function of maximizing the total Quality-Adjusted Life Years (QALYs)gained for the entire patient population over the problem horizon. Note that using this measure ensures that aggregate survival as well as quality of life of patients is considered. Although QALYs is not an operational metric that is used regularly for planning and scheduling decisions within the VHA, the management agreed that this is a reasonable objective as it is consistent with the mission of the VHA.

Calculating QALYs involves first associating each health state i with a Quality of Life (QOL) utility q^i and then multiplying the QOL utility of each health state with the corresponding number of patients in that state. These are calculated by using equations (1) through (7) developed in Section 2.4. The QOL utility is a measure of health related utility of patients and ranges between 0 and 1 where 0 corresponds to death and 1 corresponds to perfect health. Finally,the total QALYs are calculated over the entire period of analysis. Using this approach, the objective function can be represented by:

$$\sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i\left(U_{r,t}^i + W_{r,t}^i + E_{r,t}^i + M_{r,t}^i + D_{r,t}^i\right)$$

3.2 Organizational Constraints

We consider two main sources of organizational constraints in our model. The first is concerned with total annual HIV related budget at the level of a station, while the second defines service level constraints in various parts of the health care system within the station.

Budget Constraint

The budget at the GLA station consists of three components: the screening cost, health care costs associated with a patient in a particular system state and the cost of wages. This is represented by the following set of inequalities:

$$\sum_{i\in\mathcal{I}_w,r\in\mathcal{R},t\in\mathcal{M}_\tau} CS^i\alpha\beta S_{r,t}U_{r,t}^i + \sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau} CS^i U_{r,t}^i$$

$$+ \sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau,X\in\mathcal{X}} C_X^i X_{r,t}^i + \sum_{l\in\mathcal{L},k\in\mathcal{W},t\in\mathcal{M}_\tau} n_{k,l}w_k \le B(\tau) \qquad \forall\tau \qquad (8)$$

The first two terms in Equation (8) correspond to the screening costs. This is got by multiplying the cost of screening per patient in health state i (CS^i) with $\sum_{j\in\mathcal{I}_w}\alpha\beta U_{r,t}^i$, representing the asymptomatic patients

who accepted the offered HIV test and with$\sum_{j\in\mathcal{J}_o/\{13\}} U_{r,t}^{i}$ denoting the number of symptomatic HIV patients who were transferred straight to testing. Both these terms are aggregated across all risk categories and time periods up to one year. The third term represents the cost of providing healthcare services to patients in different system states. This cost is composed of several components which depend on the system state of the patient. For example, if a patient is in treatment, the cost components would be pharmacy, testing, inpatient, outpatient and overhead costs. Further, the magnitude of this component will also depend on the health state of the patients. For instance, more critically ill patients with lower CD4 count would typically incur higher pharmacy costs. We combine all such cost components into one parameter, C_X^i representing the cost of having one patient in health state $i$at system state X. Here, $X\epsilon\mathcal{X} = \{U, W, E, M, D\}$= {Unscreened, Waiting for results, Waiting to be Enrolled, Monitoring, Treatment}. The fourth term in the equation above is the labor cost which is the salary by resource type k multiplied by the staffing level of that resource type at a particular location l.

Service Level Constraints

In addition to the budget constraint, the management of the GLA station would also like to ensure timely service of patients and avoid long delays. We model this requirement using a constraint $P\{W_l \leq \tau_l\} \geq \alpha_l \ \forall l \in \mathcal{L}$ where $W_l$is the random waiting time at location l. This can be interpreted as the probability that the waiting time is less than a specified quantityτ_l, must be greater than a certain threshold α_l. Here, the tuple (τ_l, α_l) were specified at each location based on the organizational goals at the VHA. We use an M/M/1 queuing model to approximate $P\{W_l \leq \tau_l\} = 1 - e^{-(\mu_l-\lambda_l)\tau_l} \geq \alpha_l \ \forall l \in \mathcal{L}$ (Kleinrock, 1975). Here, λ_l denotes the arrival rate at location l, while μ_l denotes the service rate at location l. Using the natural logarithm operator this can be reformulated as:

$$\lambda_l \leq \mu_l + \frac{1}{\tau_l}\ln(1-\alpha_l) \qquad (9a)$$

Since the second term on the right hand side of constraint (9a) is negative, this constraint is tighter than the traditional capacity feasibility condition $\lambda_l \leq \mu_l$, which does not impose any requirements on waiting times. Note that reducingquantity$\tau_l$or increasingthresholdα_l, reduces the effective capacity $\bar{\mu}_l = \mu_l + \frac{1}{\tau_l}\ln(1-\alpha_l)$and further tightens this constraint. To operationalize (9a), we need to compute $(\lambda_l, \mu_l)\ \forall l$. The capacity of resource$k$ at location l is given by $n_{k,l}A_{k,l}\ /y_{k,l}$ patients. Therefore, we approximate the service rate at location $l$as the minimum or bottleneck capacity across all the resource or worker types available at that location given by$\mu_l = min_k\{n_{k,l}A_{k,l}\ /y_{k,l}\}$.Below we use the system dynamics developed in Section 2.4 to calculate $\lambda_l$and derive the service level constraints for each location.

Primary Care: ($l = P$). Observe from Figure 1 that the number of patients to be screened in period t is given by $\sum_{i\in\mathcal{I}_W, r\in\mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\}, r\in\mathcal{R}} U^i_{r,t}$. Therefore, $\lambda_P = \sum_{i\in\mathcal{I}_W, r\in\mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\}, r\in\mathcal{R}} U^i_{r,t}$ and $\mu_P = min_k\{n_{k,P} A_{k,P} / y_{k,P}\}$. Substituting these in inequality (9a), we get the service level constraint for screening as:

$$\sum_{i\in\mathcal{I}_W, r\in\mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\}, r\in\mathcal{R}} U^i_{r,t} \leq min_k\{n_{k,P} A_{k,P} / y_{k,P}\} + \frac{1}{\tau_P}\ln(1-\alpha_P) \quad \forall t \qquad (9)$$

Laboratory:($l = L$). Figure 1 shows that the number of patients who receive their results is $\sum_{i\in\mathcal{I}/\{13\}, r\in\mathcal{R}} R^i_{r,t}$, which is also the input rate, under the assumption of stability Therefore, $\lambda_L = \sum_{i\in\mathcal{I}/\{13\}, r\in\mathcal{R}} R^i_{r,t}$ and $\mu_L = min_k\{n_{k,L} A_{k,L} / y_{k,L}\}$. Substituting these in inequality (9a), we get the service level constraint for laboratory as:

$$\sum_{i\in\mathcal{I}/\{13\}, r\in\mathcal{R}} R^i_{r,t} \leq min_k\{n_{k,L} A_{k,L} / y_{k,L}\} + \frac{1}{\tau_L}\ln(1-\alpha_L) \quad \forall t \qquad (10)$$

Specialty Care: ($l = S$). In each period there are two kinds of patients who visit the infectious diseases specialty, patients under monitoring and patients under treatmentgiven by $M^i_{r,t}$ and $D^i_{r,t}$ respectively. Patients of health state i who are under monitoring and treatment visit the healthcare system during a given period with frequency φ^i_M and φ^i_D respectively.Therefore, $\lambda_S = \sum_{i\in\mathcal{I}/\{13\}, r\in\mathcal{R}} (M^i_{r,t}\varphi^i_M + D^i_{r,t}\varphi^i_D)$ and $\mu_S = min_k\{n_{k,S} A_{k,S} / y_{k,S}\}$. Substituting these in inequality (9a), we get the service level constraint at the infectious diseases specialty as:

$$\sum_{i\in\mathcal{I}/\{13\}, r\in\mathcal{R}} (M^i_{r,t}\varphi^i_M + D^i_{r,t}\varphi^i_D) \leq min_k\left\{n_{k,S} A_{k,S} / y_{k,S}\right\} + \frac{1}{\tau_S}\ln(1-\alpha_S) \quad \forall t \qquad (11)$$

3.3 Planning Problem

Using the above described objective function, system dynamics and organizational constraints, the planning problem faced by the GLA station can be formulated as the following nonlinear mixed integer program, which we describe as the QALY Maximizing Planning Problem (QMPP).

(QMPP) $\quad Maximize \sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i\left(U^i_{r,t}+W^i_{r,t}+E^i_{r,t}+M^i_{r,t}+D^i_{r,t}\right)$

Subject to: (1) through (11) and

$$0 \le S_{r,t} \le 1 \qquad \forall r,t \qquad (12)$$

$$U^i_{r,t}, W^i_{r,t}, R^i_{r,t}, E^i_{r,t}, M^i_{r,t}, D^i_{r,t}, ID^i_{r,t}, I^i_{r,t}, IM^i_{r,t} \in \mathbb{R}_+ \qquad \forall r,i,t \qquad (13)$$

$$n_{k,l} \in \mathbb{N}_+ \qquad \forall k,l \qquad (14)$$

Here, as developed in Section 2.4,constraints (1) through (7) describe the system dynamics.As described in Section 3.2, constraints (8) represent the budgetary constraints, while constraints (9) through (11) represent the service level constraints. Constraints (12) represent the range for the screening variable while constraints (13) and (14) represent the domains for the other variables.

Observe that the QMPP contains a knapsack problem defined by constraints (8). Thus, we need to solve instances of a NP-complete problem and it may not be always possible to solve real sized problems to optimality. We verified this in our computational experiments in Section 5. Consequently, to solve this problem, we elected to develop effective heuristics that are both computationally tractable and managerially intuitive. We also develop relaxations to the problem to obtain an upper bound on the objective function, which is used to evaluate the performance of the heuristics. If we replace $\alpha\beta S_{r,t}U^i_{r,t}$with $V^i_{r,t}$ in constraints (1), (2), (8) and (9) of the QMPP and add the definitional constraint $V^i_{r,t} = \alpha\beta S_{r,t}U^i_{r,t}, \forall r,i,t$, then the QMPP can be transformed into the following integer bilinear program QMPPB. This will be useful in developing a tight upper bound for the QMPP.

(QMPPB) $\quad Maximize \sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i\left(U^i_{r,t}+W^i_{r,t}+E^i_{r,t}+M^i_{r,t}+D^i_{r,t}\right)$

Subject to: (3) through (7), (10) through (14) and

$$U^i_{r,t+1} = \left(\sum_{j\in\mathcal{I}_\mathcal{W}} \theta^{j,i}_{r,untreat}\left(U^i_{r,t}-V^i_{r,t}\right)\right) + N^i_{r,t+1} + R^0_{r,t}\theta^{0,i}_{r,untreat} \qquad \forall r,i,t \qquad (1')$$

$$W^i_{r,t+1} = \sum_{j\in\mathcal{I}} W^j_{r,t}\theta^{j,i}_{r,untreat} + \sum_{j\in\mathcal{I}_\mathcal{W}} V^j_{r,t}\theta^{j,i}_{r,untreat} + \sum_{j\in\mathcal{I}} U^j_{r,t}\theta^{j,i}_{r,untreat} - \sum_{j\in\mathcal{I}} R^j_{r,t}\theta^{j,i}_{r,untreat} \qquad \forall r,i,t \qquad (2')$$

$$\sum_{i\in\mathcal{I}_W, r\in\mathcal{R}, t\in\mathcal{M}_\tau} CS^i V^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\}, r\in\mathcal{R}, t\in\mathcal{M}_\tau} CS^i U^i_{r,t} + \sum_{i\in\mathcal{I}/\{13\}, r\in\mathcal{R}, t\in\mathcal{M}_\tau, X\in\mathcal{X}} C^i_X X^i_{r,t} \qquad \forall\tau \qquad (8')$$

$$+ \sum_{l\in\mathcal{L}, k\in\mathcal{W}, t\in\mathcal{M}_\tau} n_{k,l} w_k \le B(\tau)$$

$$\sum_{i\in\mathcal{I}_W, r\in\mathcal{R}} V^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\}, r\in\mathcal{R}} U^i_{r,t} \le min_k\{n_{k,P} A_{k,P} \,/ y_{k,P}\} + \frac{1}{\tau_P}\ln(1-\alpha_P) \qquad \forall r,i,t \qquad (9')$$

$$V^i_{r,t} = \alpha\beta S_{r,t} U^i_{r,t} \qquad \forall r,i,t \qquad (15)$$

$$V^i_{r,t} \in \mathbb{R}_+ \qquad \forall r,i,t \qquad (16)$$

Observe that in the integer bilinear program QMPPB, all the non-linearity in the problem is now captured by bilinear constraints (15).

Proposition 1: The objective function of the QMPPB can be written as:

$$K_0 + \sum_{i\in\mathcal{I}, r\in\mathcal{R}, t\in\mathcal{M}_\tau, \tau\in[T]} \pi^i_{r,t} D^i_{r,t}$$

Where K_0 and $\pi^i_{r,t} = f(\theta^{ji}_{r,treat}, \theta^{ji}_{r,untreat}, q^i, t)$ are constants.

All proofs are provided in the Electronic Companion. Proposition 1 implies that the QALYs in the system cannot be maximized by increasing the screening rate alone as advocated by both the risk based and routine screening policies unless that increase can be translated to patients treated. This is consistent with observations in population level studies (Long et al. 2010). However, the number of patients treated is often constrained by the budgetary and capacity constraints. Thus, the focus should be on determining how many patients can be optimally treated and this in turn should be used to determine the screening rates. This is accomplished by the QMPPB. Let $\underline{U}^i_{r,t}$ be a lower bound and $\overline{U}^i_{r,t}$ be an upper bound on $U^i_{r,t}$. The computations of these bounds are described in the Appendix. The following proposition helps in reducing the complexity of the search space for heuristics to solve the QMPPB.

Proposition 2: The screening rate is bounded by the following two inequalities:

$$\sum_{r\in\mathcal{R},t\in\mathcal{M}_\tau}\sigma_{r,t}S_{r,t}\le B(\tau)-K_\tau-\sum_{i\in\mathcal{I}/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}\rho^i D^i_{r,t}\quad\forall\tau$$

$$\sum_{i\in\mathcal{I}_W,r\in\mathcal{R},t\in\mathcal{M}_\tau}\alpha\beta S_{r,t}\overline{U}^i_{r,t}\ge\sum_{i\in\mathcal{I}/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}D^i_{r,t}-\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau}\overline{U}^i_{r,t}$$

Where,K_τ, ρ^i and $\sigma_{r,t}$ are :

$$K_\tau=\sum_{k\in\mathcal{W},t\in\mathcal{M}_\tau}\left\{\sum_{i\in\mathcal{I}_o}\left(\frac{w_k y_k}{A_{k,P}}\right)\underline{U}^i_{r,t}-\left(\frac{w_k y_k}{A_{k,P}\tau_P}\right)ln(1-\alpha_P)-\left(\frac{w_k y_k}{A_{k,S}\tau_S}\right)ln(1-\alpha_S)\right\}$$

$$+\sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau}C^i_U\underline{U}^i_{r,t}+\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^i\underline{U}^i_{r,t}$$

$$\rho^i=C^i_D+\sum_{k\in\mathcal{W}}\left(\frac{w_k y_k}{A_{k,P}\tau_P}\right)\varphi^i_D$$

$$\sigma_{r,t}=\sum_{i\in\mathcal{I}_W}\left\{\sum_{k\in\mathcal{W}}\left(\frac{w_k y_k}{A_{k,P}}\right)\underline{U}^i_{r,t}+CS^i\underline{U}^i_{r,t}\right\}$$

Further, for a stationary screening policy for which $S_{r,t}=S_r\ \forall t$, $S_r\le\frac{B(\tau)-K_\tau}{\sum_{t\in\mathcal{M}_\tau}\sigma_{r,t}}$.

Note from Proposition 2 that for a given screening rate, the total number of patients that can be treated is bounded by: 1)The residual budget left over for treatment after the screening, staffing and the patient state costs. 2) The number of screened asymptomatic patients who test positive and symptomatic patients being treated. Further, the total number of patients who actually are treated will be determined by whichever of these two conditions become tight. Given, that typically budgets are scarce and there is a large population of patients, it is likely that the budget constraint would be tighter. This implies that while setting screening rates, one has to understand budgets and its implications on treatment. This is consistent with the public health literature (Martin et al. 2010).

3.4 Relaxations and Upper Bounds

To develop an upper bound on the QMPPB, we replace bilinear constraints (15) by convex over and under estimators of the bilinear terms using the approach proposed by McCormick (1976).

Let$\overline{U}_{r,t}^{i}$ and$\underline{U}_{r,t}^{i}$ represent the upper and lower bound on the variable $U_{r,t}^{i}$respectively. Then it follows from (15) that:

$$V_{r,t}^{i} \geq \alpha\beta S_{r,t}\underline{U}_{r,t}^{i} \qquad \forall r,i,t \qquad (15a)$$

$$V_{r,t}^{i} \leq \alpha\beta S_{r,t}\overline{U}_{r,t}^{i} \qquad \forall r,i,t \qquad (15b)$$

Note that $\alpha\beta\ \underline{U}_{r,t}^{i} \leq \alpha\beta U_{r,t}^{i} \leq \alpha\beta\overline{U}_{r,t}^{i}$and $0 \leq S_{r,t} \leq 1 \forall r,i,t$. Then, $\alpha\beta S_{r,t}\overline{U}_{r,t}^{i} + \alpha\beta U_{r,t}^{i} - \alpha\beta\overline{U}_{r,t}^{i} = (S_{r,t}-1)\alpha\beta\overline{U}_{r,t}^{i} + \alpha\beta U_{r,t}^{i} \leq (S_{r,t}-1)\alpha\beta U_{r,t}^{i} + \alpha\beta U_{r,t}^{i} = S_{r,t}\alpha\beta U_{r,t}^{i} = V_{r,t}^{i}$. Thus,

$$V_{r,t}^{i} \geq \alpha\beta S_{r,t}\overline{U}_{r,t}^{i} + \alpha\beta U_{r,t}^{i} - \alpha\beta\overline{U}_{r,t}^{i} \qquad \forall r,i,t \qquad (15c)$$

Similarly, $\alpha\beta S_{r,t}\underline{U}_{r,t}^{i} + \alpha\beta U_{r,t}^{i} - \alpha\beta\underline{U}_{r,t}^{i} = (S_{r,t}-1)\alpha\beta\underline{U}_{r,t}^{i} + \alpha\beta U_{r,t}^{i} \geq (S_{r,t}-1)\alpha\beta U_{r,t}^{i} + \alpha\beta U_{r,t}^{i} = S_{r,t}\alpha\beta U_{r,t}^{i} = V_{r,t}^{i}$. Thus,

$$V_{r,t}^{i} \leq \alpha\beta S_{r,t}\underline{U}_{r,t}^{i} + \alpha\beta U_{r,t}^{i} - \alpha\beta\ \underline{U}_{r,t}^{i} \qquad \forall r,i,t \qquad (15d)$$

Observe thatconstraints (15a) through (15d) provide a linear relaxation to bilinear constraints (15). This substitution reduces this problem to a linear mixed integer program which can now be solved to optimality using commercial solver such as the GUROBI solver (Gurobi Inc., 2010). We call this formulation as the RQMPPB and note that the optimal solution to the RQMPPB provides an upper bound to the QMPPB and by Proposition 2, to the QMPP.

The quality of this upper bound strongly depends on the bounds of $U_{r,t}^{i}$. A recent improvement to the McCormick relaxation is introduced by Wicaksono and Karimi (2008). We adapt this technique to do an *ab initio* partitioning on $U_{r,t}^{i}$, apply a set of under and over estimators to each partition and introduce a logical constraint to limit the partitioned variable to one active partition. To achieve this, let $U_{r,t}^{i}$ be separated into $m$equally spaced partitions as $\underline{U}_{r,t}^{i} = a_{r,t}^{i}(1) < \ldots < a_{r,t}^{i}(m) < a_{r,t}^{i}(m+1) = \overline{U}_{r,t}^{i}$. The choice of parameter$m$ is based on comparing the reduction in the value of the bound with the increased time it takes to compute the bound when m is incremented by one starting with $m = 1$and is described in the Electronic Companion. Define binary variable $\xi_{r,t}^{i}(m)$ so that $\xi_{r,t}^{i}(m) = 1$ if $U_{r,t}^{i} \in [a_{r,t}^{i}(m), a_{r,t}^{i}(m+1)]$ and $\xi_{r,t}^{i}(m) = 0$ otherwise. This leads to the following constraints:

$$U_{r,t}^{i} \geq a_{r,t}^{i}(m)\xi_{r,t}^{i}(m) + \underline{U}_{r,t}^{i}\left[1 - \xi_{r,t}^{i}(m)\right] \qquad \forall\, r,i,t,m \qquad (15e)$$

$$U_{r,t}^{i} \leq a_{r,t}^{i}(m+1)\xi_{r,t}^{i}(m) + \overline{U}_{r,t}^{i}\left[1 - \xi_{r,t}^{i}(m)\right] \qquad \forall\, r,i,t,m \qquad (15f)$$

$\sum_{m=1}^{M} \xi_{r,t}^{i}(m) = 1 \quad \forall\, r, i, t$ (15g)

$\xi_{r,t}^{i}(m) \in \{0,1\} \quad \forall\, r, i, t, m$ (15h)

Next, we introduce constraints of the type (15a) through (15d) for each partition by replacing $\overline{U}_{r,t}^{i}$ with $a_{r,t}^{i}(m+1)$ and $\underline{U}_{r,t}^{i}$ with $a_{r,t}^{i}(m)$. Depending on $\xi_{r,t}^{i}(m)$, the appropriate set of constraints would be activated, thus providing tight relaxation to the bilinear terms. This leads to the following constraints:

$V_{r,t}^{i} \geq \alpha\beta S_{r,t} a_{r,t}^{i}(m) - K\left[1 - \xi_{r,t}^{i}(m)\right] \qquad \forall\, r, i, t, m$ (15a')

$V_{r,t}^{i} \leq \alpha\beta S_{r,t} a_{r,t}^{i}(m+1) + K\left[1 - \xi_{r,t}^{i}(m)\right] \qquad \forall\, r, i, t, m$ (15b')

$V_{r,t}^{i} \geq \alpha\beta S_{r,t} a_{r,t}^{i}(m+1) + \alpha\beta U_{r,t}^{i} - \alpha\beta a_{r,t}^{i}(m+1) - K\left[1 - \xi_{r,t}^{i}(m)\right] \qquad \forall\, r, i, t, m$ (15c')

$V_{r,t}^{i} \leq \alpha\beta S_{r,t} a_{r,t}^{i}(m) + \alpha\beta U_{r,t}^{i} \quad - \alpha\beta \quad a_{r,t}^{i}(m) + K\left[1 - \xi_{r,t}^{i}(m)\right] \qquad \forall\, r, i, t, m$ (15d')

The value of parameter K is set sufficiently large to deactivate these constraints if $U_{r,t}^{i}$ does not belong to that particular partition. To provide a tighter upper bound on the QMPPB, we solve the RQMPPB by replacing (15a) through (15d) with (15a') through (15d') and (15e) through (15h). The performance of this bound is evaluated in Section 5.

3.5 Heuristics and Lower Bounds

In this section, we discuss several possible heuristic solution methods to the QMPPB that correspond to potential implementation policies at the GLA station. They can broadly be classified as fixed staffing heuristics and variable staffing heuristics.

Fixed Staffing Heuristics

Here, we do not optimize over the staffing variables $n_{k,l}\ \forall k, l$ and these are set to existing levels corresponding to the risk based screening policy. In this case, QMPPB reduces to a continuous bilinear program. We then develop two heuristics depending on how the screening rate varies over time. In the first heuristic, we add constraints $S_{r,t} = S_r \quad \forall r, t$ to ensure that the recommended screening policy is stationary. Although apparently restrictive, it is easy to implement and was appealing to the GLA station management. To solve the resulting problem we iteratively narrow down on the optimal stationary fixed screening using the search algorithm described in the Electronic Companion. Note that this algorithm is quite simple to implement as evaluation of the QMPPB given the screening rates is now a linear program and can be solved very effectively using several commercially available solvers such as the GUROBI solver. Further, Proposition 2 enables us to reduce the solution space of this algorithm. We refer to this heuristic as the Fixed Staffing Stationary Screening (FSSS) heuristic.

In the second heuristic, we allow the screening rate to vary over time so that the resulting screening policy is non-stationary. The resulting problem reduces to a continuous bilinear program which is solved by using the generalized reduced gradient algorithm (Abadie and Carpenter, 1969). This algorithm has been shown to be very effective for large sparse dynamic nonlinear optimization problems (Drud, 1985). We refer to this heuristic as the Fixed Staffing Non-stationary Screening (FSNS) heuristic. Clearly this heuristic is less restrictive than the FSSS and hence can be expected to perform better. We verify this in Section 5.

Variable Staffing Heuristics

Next, we describe two heuristics, where we allow the staffing levels to change andagain consider either stationary or non-stationary screening rates. We refer to these as the Variable Staffing Stationary Screening (VSSS) and the Variable Staffing Non-stationary Screening (VSNS) heuristic respectively. The solution procedure for the VSSS heuristic is very similar to that of the FSSS heuristic, with the key difference being that the evaluation of the QMPPB for a given screening rate in the search algorithm would now require solving a mixed integer program. While this potentially can be more complicated, we found that the GUROBI solved this problem very effectively. The solution to the VSNS heuristic is complicated as it involves solving a nonlinear mixed integer program. We employ the combined penalty and outer approximation method (Vishvanathan and Grosman, 1990) to solve this problem. Given that we can optimize both staffing levels and the screening rates in the variable staffing heuristics, we expect both of them to outperform the corresponding fixed staffing heuristics. However, the magnitude of the gap between these heuristics is not apparent. Similarly,whether the VSSS outperforms the FSNS or vice versa is not obvious a priori. We investigate these issues in the computational experiments in Section 5.

Finally, observe that the QMPPB is not jointly convex in the decision variables. Thus, this sequential approach in the FSSS and the VSSS provides a feasible but not necessarily an optimal solution. Similarly, given the complexity of the QMPPB, the algorithms used to execute the FSNS and VSNS provide feasible but not optimal solutions.

4. Data Collection and Model Validation

The data required for our model can be divided into two broad categories. The first category includes operational data concerning costs, budgets, incoming patient characteristics, time required for various activities, time available and service level parameters. These data are specific to the GLA station and were collected from a variety of sources including direct observation, administrative databases, clinical studies and discussions with the GLA stationmanagement. The second category includes clinical data on visit frequency under HIV care, the quality of life estimates for HIV patients in different health statesand on treatment decisions. Weused published estimates for these parameters from the existing clinical

literature and are more broadly applicable. Below we describe each of these categories in greater detail. We then use the data to validate our model both in the context of the literature and the GLA station.

4.1 Operational Data

Costs

Primary drivers for variable cost in our model are cost of HIV screening cost (CS^i), system state cost (C_X^i) per patient and wages (w_k).The screening cost CS^i consists of the material cost of screening. The screening cost per patient was estimated to be $80 after discussions with the management. The system state cost per patient C_X^i is composed of several components. Therefore, its estimation is more involved and discussed in the Electronic Companion. As the staffing levels are endogenous to the model, the other relevant cost component are the wagespaid to the health care workers of different types (w_k). At the GLA station, these costs are fixed and do not vary based on the patient load. These are shown in Table 2A in the Electronic Companion.

Budget

The VHA allocates the budget to the GLA station annually and this budget does not carry over to the next year. To provide a more stable and a long range plan, the management at the GLA station suggested that we conduct our analysis for a period of 2 years, where the budget for year τ is given by $B(\tau)$, $\tau\epsilon\{1,2\}$. Note that, our model can be easily extended for $\tau > 2$ without any changes to the methodology by the appropriate choice of T, where $\tau \in [T] = \{1, 2, \ldots, T\}$. This is described in the Electronic Companion. However, extending the model beyond two years was not realistic in our application context as there was significant uncertainty in the costs of screening and treatment, the population of veterans that would be served at this station, and the incidence and prevalence rates. To incorporate the uncertainty in these parameters, the model can be solved every year with a two year horizon usingupdated parameters.

Due to various complexities in estimation, the annual GLA station budget was not broken down to the level for HIV related activities which is the focus of our analysis. Therefore, we imputed a budgetary range $\left[\underline{B}(\tau), \overline{B}(\tau)\right]$using the risk based screening policy currently followed at VHA (i.e.,$S_{1,t} = 1 \,\&\, S_{2,t} = 0 \;\forall t$). The lower bound of this range corresponds to the smallest annual budget at which the risk based screening policy is feasible. The upper bound corresponds to the smallest value of the annual budget at which no further gains in QALYS can be accrued from the risk based screening policy. This approach to calculate $\left[\underline{B}(\tau), \overline{B}(\tau)\right]$ is formalized in the budget imputation algorithm provided in the Electronic Companion. We conduct our analysis on all the proposed policies in Section 3.5 within this budgetary range.

Incoming Patient Characteristics

LetN_t denote the number of new patients entering the station in time period t and $\hat{p}_r^i$ be the fraction of these patients in risk category r and health state i. The number of new patients in each risk category and health state in each period who enter the station is thus given by$N_{r,t}^i = N_t \hat{p}_r^i$. To estimate N_t we calculated the mean of historical data of total incoming patients over the past 12 months. The variation around the mean was negligible and we did not detect any temporal trends (such as increasing or decreasing over time) for the number of new patients.The parameter $\hat{p}_r^i$ is the proportion of patients in each risk and CD4 category. We calculate $\hat{p}_r^0 = (1 - prev_r)$, where prevalence rate ($prev_r$) is estimated by Paltiel et al. (2005) and shown in Table 3A in the Electronic Companion.The proportion of patients who are infected ($prev_r$) are further divided into different CD4 counts in a fraction estimated for the VHA by Gandhi et al. (2007), thus determining $\hat{p}_r^i, \ \forall i \neq 0$. Wereport this in Table 4A in the Electronic Companion. Themanagementprovided us with U_1, the total number of patients currently enrolled at the GLA station. Thus, the number of unscreened patients in each risk category and each health state would be given by$U_{r,1}^i = \hat{p}_r^i \ \ U_1$.

The fraction of patients who visit a health care facility for non-HIV related reasons $\alpha$was estimated by dividing the total number of unique patients who visited the inpatient or the outpatient facilities for non-HIV related reasons by the total number of patients registered in the station. Using this approach, we estimated $\alpha = 0.5$. The proportion of patients who accept screening β was assumed to be 50% based on prior studies (Goetz et al. 2008b).

Time Required, Time Available and Service Level Parameters

To estimate$y_{k,l}$, the time required per patient of health care worker of type k at location l, we used an observational time and motion study conducted in the emergency department in the west Los Angeles Veterans hospital within the GLA station (Gidwani et al. 2009). This data shown in Table 5A in the Electronic Companionwas validated against other published estimates (Silva et al. 2007). Further, in our discussion, the GLA station management noted that these times would be very similar for other care settings in the station such as the primary care clinics, inpatient and outpatient departments.

The total time available at each resource at each location per month, $A_{k,l}$ for activities associated with the routine HIV screening program was based on the discussion with the GLA station management. It took into account the fact that health care workers need to devote time to other clinical and administrative activities as well. These estimates are shown in Table 6A in the Electronic Companion.

Lastly, the management at the GLA station provided the service level requirements at each location. It was expected that at least95% of all patients should be processed at each location within a period of one month. Thus, $\tau_l = 1, \alpha_l = 0.95$.

4.2 Clinical Data

Visit Frequency under HIV Care

The outpatient visit frequency for VHA was not directly available. We used published estimates by Schackman et al.(2006) for the frequency of outpatient visit under monitoring (φ_M^i) and under treatment (φ_D^i). This is reported in Table 7A in the Electronic Companion.

Quality of Life (QOL)Utilities

The QOL utilitiesweredrawn from Freedberg et al. (1998) and Mauskopf et al. (2005). These are summarized in Table 3 and more details are provided in the Electronic Companion.Here, it was assumed that the health related quality of life utilities (q^i) are directly associated with the underlying health state represented by the CD4 count category and OI infection status rather than on the treatment status, per se. This is reasonable because the effect of treatment is eventually reflected in patients being in better health states and hence enjoying a higher QOL utility.

[Insert Table 3 here]

Treatment Decision

The treatment policy at the GLA station was to initiatepatients having CD4 cell count below 350 cells/mm^{3}and patients with opportunistic infection irrespective of their CD4 count ontreatment and retain the reston monitoring.From Table 1, this implies that $z^i = 0$ for $i = \{0, 1, 2\}$and $z^i = 1$, otherwise.

4.3 Model Validation

In this section, we conduct analyses to validate the model in the context of the literature and the GLA station. To ensure an unbiased comparison with the literature (Paltiel et al. 2005; Bishai et al. 2007), we removed all the organizational constraints in the model so that it reduces to a pure disease progression and treatment model as considered by these papers. Bishai et al. (2007) calculate total QALYs gained from treatment over no treatment for HIV positive patients. We used their treatment regimen in our model and found that the total QALYs gained was comparable to their work. Paltiel et al. (2005) calculates the $ spent per QALY gained from going from no treatment to treatment under various screening policies and found that this varied between $63,000 and $113,000 spent per QALY gained. We also used our model to calculate the $ spent per QALY gained for the different policies in Paltiel et al. (2005) and found it to be similar, ranging from $61,000 to $111,000 spent per QALY gained. This validates that our disease progression and treatment model is consistent with the literature.

In the context of the GLA station, we considered the entire model and the current risk based screening policy. We found that the model estimates on the number of people at each disease state, location and time period were within 2% of the actual numbers at the GLA station. We also used the resulting arrival rate λ_l and service rate μ_l at location $l \in \{P, L, S\}$to estimate $\overline{W}_l = 1/(\mu_l - \lambda_l)$, the average wait times at each location for a given time period under the M/M/1 queuing model assumption used in deriving the

service level constraints (Kleinrock, 1975). We found these estimates were within 5% of the actual average wait times for the corresponding locations and time period at the GLA station. This supported the rationale for using the M/M/1 queuing model in developing the service level constraints. These analyses also validate that our model effectively captures the operating environment at the GLA station and is a necessary step to provide confidence in the policy analysis described next.

5. Policy Analysis

In this section, we evaluate several policies for screening, testing and care within the framework of our model. We start with analyzing the risk based screening policy which had been the standard of care at the VHA when we started our collaboration. We then evaluate the impact of the routine screening policy under consideration and also assess the performance of the heuristics described in Section 3.5.

Recollect from Section 4.1 that the annual budget expenditure required for HIV screening, treatment and monitoring was not directly available. Therefore, we used the budget imputation algorithm provided in the Electronic Companion to first to impute the budget range $[\underline{B}(\tau), \overline{B}(\tau)]$for the risk based screening policy in which $S_{1,t} = 1 \,\&S_{2,t} = 0 \;\forall t$. Here, we found that $\underline{B}(\tau) = \$10$ million and $\overline{B}(\tau) = \$20$ million for $\tau = 1$ and 2. This implies that at least \$10 million is needed annually to implement the risk based screening program and any budget allocation over \$20 million will not improve the efficacy of this program further. We also used this algorithm to find that an annual budget of \$35 million was required to implement the routine screening policy in which $S_{r,t} = 1 \;\forall r, t$. The management at the GLA station found this input instructive but also felt that this level of funding would not be available in the foreseeable future. Instead, they were interested in improving upon the risk based policy but within the current budgetary range of \$10 to \$20 million. To perform this analysis and simplify the exposition, we conducted all our subsequent analysis at three budget levels: low, medium and high corresponding to \$14, \$16 and \$19 million respectively. These values are chosen in discussion with the station management. We tried to solve the QMPP for these budget values using leading commercial solversfor nonlinear mixed integer programs such as BARON and DICOPT using the NEOS server (Dolan 2002). However, in all cases, these solvers could not even generate feasible solutions after over forty hours of computation and the runs were aborted. This provides validation for developing boundsand heuristics to address this problem.

5.1 Performance of Heuristic Policies

We solved the FSSS, FSNS, VSSS and VSNS using the approaches described in Section 3.5 and then calculated the QALYs gained from these four heuristic policies. We used the technique described in Section 3.4 to compute the upper bounds for each of these budgetary levels. The computations for the risk

based screening policy, the routine screening policy, FSSS, VSSS and upper bounds were executed with GUROBI, a general purpose LP/MIP solver using the NEOS server. The computations for the FSNS and the VSNS were implemented with DICOPT using the NEOS server. All heuristics were solved in a few seconds while each computation of the upper bound took at most three hours. Note that in computing the upper bounds for the fixed staffing heuristics FSSS and FSNS, we fixed the staffing levels at the current levels at the GLA station. This ensured that these heuristics were being fairly compared to an upper bound to the fixed staffing problem. We measured the performance of the heuristics using % gap defined as the difference between QALYs gained from the upper bound and those gained from the heuristic policy expressed as a percentage of the QALYS gained from the upper bound. In all cases, QALYs gained were calculated with the base case of no screening. Table 4 summarizes the gaps for the four heuristics across the three budgetary levels.

[Insert Table 4 here]

The % gaps described in Table 4 indicate that all the heuristics perform very well. In particular, the average gap across these heuristics is 1.95% and ranges from 0.08% to 5.15%. In general, for the fixed staffing heuristics the gaps increase as the budget level increases. This is because the upper bounds increase at a greater rate than the heuristic solution. The rate of growth of the heuristic solution is limited as the benefits from choosing the optimal screening rates at higher budget levels saturates due to fixed staffing in which more patients cannot be treated due to capacity and service level constraints. Conversely, for the variable staffing heuristics, the gaps decline as the heuristic solution increases at a greater rate than the upper bound. This is because variable staffing allows more effective allocation of staff at the higher budget levels to treatment, allows more screened patients who are diagnosed with HIV to be treated optimally and this improves the overall performance of the heuristics.

We also conducted sensitivity analysis to understand how parameters such as time available for HIV screening programs, service level parameters and the costs of wages, screening and treatment affect these gaps for the heuristics. To perform this analysis, we first set the budget level to $16 million and changed each of these parameters one at a time from their base level by -30% to 30% in increments of 10%. We then calculated the gap for each heuristic and the appropriate change in the gap from the baseline reported in Table 4. Across all heuristics and range of values of these parameters, we found the average *change* in gaps was 3.3% and this varied from 0.8% to 7.2%. This shows that these heuristics and the upper bounds are robust across a wide range of parameter values.

5.2 Improvements fromRisk Based Screening

We computed the QALYs accrued at these budget levels for the current risk based screening policy. We used this to calculate the % improvement of the heuristics from the risk based screening policy expressed as a percent of the risk based screening policy solution. The results, summarized in Table 4,lead to the

following observations. First, irrespective of the budget level, improvements from risk based screening increasedas we go from the FSSS to the FSNS to the VSSS and finally to the VSNS heuristic. In particular, the most improvement is obtained from the VSNS because this policy synchronizes the screening decision with the staffing decision. This is important since it is ineffective to screen as many patients as possible and not have sufficient funding to treat them as necessary. Rather, it is critical to screen as many patients that can be optimally treated as the benefits arise only from treatment and not screening. This was shown in Proposition 1. This implies that one should first calculate how many people can be optimally treated and then use this to appropriately calculate the optimal screening rates. This approach is executed by the solution method of the VSNS. Second, note that the FSNS improves upon the FSSS by at most 3.47% and this is only 0.14% in the most realistic low budget scenario. This suggests that if staffing cannot be changed due to organizational reasons, then it is better to keep a stationary screening policy in the short term, since this is easier to implement. However, if the long term goal is to accrue maximum benefit using the VSNS, the FSNS would be a good approach to allow the staff to get acclimatized to using non stationary screening rates prior to implementing the more radical changes associated with variable staffing. Third, the gains from varying staffing are more significant than those obtained by varying screening across any budget level. To see this, observe from Table 4 that the gains from going from fixed to variable staffing (i.e., FSSS to the VSSS or FSNS to the VSNS) are larger than the gains from stationary to non-stationary screening(i.e., FSSS to the FSNS or VSSS to the VSNS). Fourth, the benefit from variable screening is greater if staffing is allowed to change (i.e., the gains from VSNS-VSSS > FSNS-FSSS). Finally, the greatest improvements from current practice occur in low budgets or resource constrained environments. This is because the optimization executing these policies ensures that screening and staffing rates are chosen in such a manner that these scarce resources are used in the best possible manner.

Finally, we again conducted sensitivity analysis to study how the % improvement of the heuristics from the risk based screening policy change with model parameters such as time available for HIV screening programs, service level parameters and the costs of wages, screening and treatment. To do so, we first set the budget level to \$16 million, and changed each of these parameters one at a time from their base level by -30% to 30% in increments of 10%. In practice, such changes may be needed due to organizational requirements. As expected, the QALYs gains from all the heuristics declined as available time for HIV programs ($A_{k,l}$) and the service level parameter related wait time at location l (τ_l) decreased. Similarly, the QALYs gained from the heuristics declined as the service level parameter related to the probability of meeting a wait time at location l (α_l), cost of wages, screening and treatment increased. However, in all these cases, the *relative* gain from the benchmark risk based screening policy are increasing as the optimization inherent in the heuristics allowed them to better cope with diminished

resources or higher service level requirements or increased costs. In addition, the previously described order of improvement from FSSS to FSNS to VSSS to VSNS was still preserved. This shows that the comparative performance of the heuristics across a wide range of parameters is quite consistent and they are better in coping with changes in these parameters values than the risk based screening policy.

5.3 Screening Rates and Staffing Allocation

We studied how the screening rates and staffing allocation vary for each of these policies at different budget levels. We start by discussing the screening rates across the policies. Here, we found at low budget levels, the screening rates of the variable staffing heuristics were higher than those of the fixed staffing heuristics. This is because fixing the staffing levels to those of the risk based screening policy resulted in a large portion of the budget being committed,thereby leaving little flexibility to increase screening rates. On the other hand, at higher levels of budget, the screening rates of the fixed staffing heuristics are now higher than the variable staffing heuristics. This is because once the staffing levels are fixed, the only way to utilize the additional budget and improve the solution is to increase screening rates. In contrast, the variable staffing heuristics balances the screening rates and staffing levels with the available budget in both these budget scenarios, and thus yields a better solution. We also analyzed how screening rates vary over time in the non-stationary screening rate policies (i.e., FSNS and VSNS). Observe from Figure 2 that in both the FSNS and VSNS policies, screening rates ramp up, saturate at a stable level and ramp down across a budget horizon. The ramp up occurs as there is a large pool of unscreened patients at the start of the horizon. Screening these patients at high rates would require large number of staff at screening and thus less staff would be available at treatment. This would lead to an undesirable outcome of screening patients without treating them. To prevent this from happening, both these policies ramp up screening rates to spread the workload over time with fewer staff at screening so that the remaining staff can be effectively utilized in treatment. This ramp up continues until the system reaches the desired balance between screening and staffing and at which point, the screening rate stabilizes. This screening rate is maintained until the time horizon for the current budget cycle draws to a close. At this point, the screening rates ramp down and more resources are focused on the treatment of screened patientsto make sure that screened patients not treated in this horizon do not congest treatment in the next horizon. This is important as residual budgets from the current cycle do not carry over to the next cycle.

[Insert Figure 2 here]

Next, consider the staffing allocation between primary care (i.e., where screening is conducted) and specialty care (i.e., where treatment is conducted) across policies. This is summarized in Figure 3 and from this figure it can be seen thatmore staff was allocated to primary care compared to specialty care in the fixed staffing heuristics, whose staffing levels are set to the current risk based screening policy. This follows as in the risk based screening policy, all high risk patients are screened without explicitly

determining the staffing requirement for treatment. This leads to lower QALYs in the system as many people are screened but may not be effectively treated. Conversely, the variable staffing policies allocated more staff to specialty care than primary care. This ensured that the number of patients treated and the resulting system wide QALYs are maximized since as shown in Proposition 1, these are accrued from treatment and not from screening. Finally, we observed that the staffing level in variable staffing heuristics was actually lower than those in the fixed staffing heuristics. This was a direct consequence of optimizing the allocation between primary and specialty care in the variable staffing heuristics based on the number of patients that can be treated. This, in turn, reduced the staffing level needed at screening to a greater extent than the increase in staff needed at treatment.

[Insert Figure 3 here]

To summarize, the policy analysis conducted in this section has led to many organizational implications at the GLA station. These are discussed next.

6. Application and Discussion

Several ideas developed in this paper have influenced decision making at the GLA station. The FSSS and the FSNS have been used to compute screening rates (Anaya et al. 2012). The rates ranged from 15% to 30% for the risk categories. These rates were considered to be reasonable and achievable. Further they are consistent with researchon HIV screening rates in other health care settings (Martin et al. 2010). The rates from the FSSS and the FSNS were used to compute how many patients could be estimated to be present at the primary care, laboratory and the infectious disease specialty over time. This information was then used in constraint (8) to estimate the appropriate costs at different partsof the GLA Station. This provided valuable input for planning in future budgetary cycles. In addition our methods show how these costs changed from the risk based screening policy to the FSSS and the FSNS. This in part was useful in gaining the necessary funding in these budget cycles to implement these policies.

The implementation of the FSSS and FSNS has led to early detection and early transfer to care for an increased number of patients. This in turn, has resulted in better patient outcomes as they are identified at a stage of disease where the more serious manifestations of the illness are less common and when the response to therapy is better (Goetz and Rimland 2011). The challenges in implementing these policies include educating the patients about the procedure and benefits of early testing, overcoming the reluctance of the providers to screen and prescribe these tests to patients they considered low risk or older and in stable monogamous relationships, training the staff at primary care to execute screening correctly, ensuring tests are conducted and information passed to care in a timely manner, and ensuring that patients are connected to care in an effective manner. Once patients are connected to care, it is important that there are sufficient updates of their health state information to ensure effective planning of staff for incoming

patients in future periods. To ameliorate the impact of these challenges, the GLA station started implementation at its largest facility and used this learning to roll out to the whole station and other stations at the VHA (Goetz et al. 2011).

In addition, this work has had severalmanagerial implications. It has shown the management that even though a policy such as routine based screening may be cost effective from a societal point of view, its implementation may not be feasible in an organization due to budgetary constraints. In particular, we show that at least a $15 million or75% increase of annual budgetary outlays would be required to implement this policy from the risk based screening policy. This was not possible at the GLA station due to the existing budgetary environment. Therefore, this motivated the GLA station to improve upon the risk based screening policy and we propose the FSSS, FSNS, VSSS and the VSNS policies. Our analysis of these policies (summarized in Table 4)showed that optimizing the screening rate with existing staffing levels could increase the QALYs gained from risk based screening by 20% to 40% or to 295 and 1094 QALYs gained at the low and high budget levels respectively. Further, in the low budget scenario, optimization of screening and staffing levels could increase QALYs gained from 245 for risk based screening[5] to 995 or by over 300%. The approaches we propose improves on risk based screeningas it focuses on treatment, determines how many patients can be treated effectively and then decides the appropriate screening rate. This is crucial as treatment determines the QALYs accrued in the system. This is in contrast to risk based screening where all high risk patients are screened without consideration of the staffing implications for treatment. In particular, the staffing implications of our variable staffing policies at the GLA station are more staff should be allocated to specialty care, lesser to the primary care and this allocation in fact lowered total staff requirements. While such staffing policies are harder to implement from an organizational perspective, we show this could result in significantly more gains and this provides the management with the justification to consider these policies. Furthermore, we find that greatest benefit under variable staffing can be got by non-stationary screening. Here, it is beneficial to initially ramp up the screening rate to even the workload over time at treatment, allow this rate to stabilize and finally ramp down towards the end of the budget cycle so that the remaining budget can be effectively used for treatment of patients. Finally, it is encouraging to note that the greatest gains can be achieved by these policies from risk based screeningat the most realistic low budget scenario. In addition, the gains are increasingin order of FSSS to FSNS to VSSS to VSNS and this is independent of any budget scenario.Therefore our analysis provides direct justification for the GLA station to next consider the variable staffing policies (i.e., the VSSS and the VSNS) as the logical extension of the FSSS and the FSNS that have been adapted due to our work.Further,our method provides close to optimal staffing allocation and screening rates to successfully execute such variable staffing policies.

[5]This is consistent with the gains by risk based screening in other studies (Paltiel et al. 2005).

This work has the following limitations. First, our model does not account for the societal benefits of early screening by reducing transmission and ultimately prevalence rates. However, it is not possible to *analytically* estimate this reduction as it depends on individual behavior (i.e., whether one would take adequate precautions after being diagnosed) and if the people affected by this individual are a part of the VHA system.Therefore, we systematically reduced prevalence rates to calculate the impact on budgets and QALYs gained. The results summarized in the Electronic Companion shows that even small reduction in prevalence rates could significantly lower budget requirements or increase QALYs gained. Second, we have assumed only two risk categories in determining screening rates and do not further stratify based on race and ethnicity as there are no clinical studies which can be then used to estimate transition rates between several health states. However, such divisions may increase the efficacy of our methods by early identification and treatment of certain patient groups. Third, several model parameters such as visit frequency, QOL utilities, incidence and prevalence rates were estimated using clinical literature based on the general HIV population as they were not available specifically to the GLA station. This implies to improve the performance of our methods, these parameters need to be updated as results from more current clinical studies become available or studies specific to the GLA station are conducted. Finally, our analysis is conducted at the station level as requested by the GLA management for budgetary and staff allocation reasons. To keep this aggregate analysis tractable, we assumed a compartmental model with deterministic transitions between health states. However, this approach leads to a loss of granularity in terms of patient flows. Specifically, we do not consider the differences in cost and treatment effectiveness of individual patients in a particular health state. Further, we do not incorporate prioritization decisions that may be made within a health state due to presence of other health conditions of the patients such as heart disease, diabetes or cancer. To consider these aspects in a shorter time horizon, one needs to consider a more detailed scheduling model with stochastic transition between disease states and this is beyond the scope of our study.

In conclusion, we developeda model to address the screening and staffing decisions for HIVscreening, testing and care at the GLA station of the VHA. We applied this model to evaluate the risk based screening policythat was being used and also showed that the cost effective routine screening policy recommended by the CDC is not feasible in this organizational context due to budgetary constraints. Therefore, we developed alternate policies within the framework of our model that are feasible and determined the relative improvement from using these policies from the risk based screening policy. Two of these policies, the FSSS and FSNS are currently being used at the GLA station. We also developed managerial insights to better understand these policies and provided justification to the station administration to further extend and enhance their use by considering the variable staffing policies VSSS and the VSNS. This paper opens up several opportunities for future work. First, further work could be

done to improve the heuristic policies and the upper bound to reduce the sub optimality gap. Second, this framework can also be used to evaluate HIV screening, testing and care in other healthcare systems that hasperiodic patient follow up and in which residual budgets do not carry over to future periods (Petersen et al. 2007). In these settings, our existing modeling framework may have to be changed to include alternateobjective functions, system dynamics and organizational constraints. This could require development of different solution methods and bounds. Finally, a similar modeling framework can be used to assess the feasibility of other cost effective interventions (such as intuberculosis and cardiac care) and if needed, develop alternate policies that improve current practice and are feasible from an organizational perspective.

Appendix

Estimation of Boundson $U^i_{r,t}$

We describe the calculation of the lower bound $\underline{U}^i_{r,t}$and the upper bound $\overline{U}^i_{r,t}$on $U^i_{r,t}$. These parameters are used in Proposition 2 to reduce the search space of the search algorithms and also important parameters in the method described in Section 3.4 used to develop upper bounds on the QMPPB. From Equation (1), we get

$$U^i_{r,t+1} = \left(\sum_{j\in\mathcal{J}_W} \theta^{j,i}_{r,untreat}\left(1-\alpha\beta S_{r,t}\right)U^j_{r,t}\right) + N^i_{r,t+1} + R^0_{r,t}\theta^{0,i}_{r,untreat}$$

$$U^i_{r,t+1} \geq \left(\sum_{j\in\mathcal{J}_w} \theta^{j,i}_{r,untreat}\left(1-\alpha\beta S_{r,t}\right)U^j_{r,t}\right) + N^i_{r,t+1}$$

$$U^i_{r,t+1} \geq \left(\sum_{j\in\mathcal{J}_w} \theta^{j,i}_{r,untreat}\left(1-\alpha\beta\right)U^j_{r,t}\right) + N^i_{r,t+1}\,, since\ S_{r,t} \leq 1$$

If we can find, an$\underline{U}^i_{r,t} \leq U^i_{r,t}$, then,

$$U^i_{r,t+1} \geq \left(\sum_{j\in\mathcal{J}_w} \theta^{j,i}_{r,untreat}\left(1-\alpha\beta\right)\underline{U}^i_{r,t}\right) + N^i_{r,t+1}$$

Therefore, we get the recursive formula:

$$\underline{U}^i_{r,t+1} = \left(\sum_{j\in\mathcal{J}_w} \theta^{j,i}_{r,untreat}\left(1-\alpha\beta\right)\underline{U}^i_{r,t}\right) + N^i_{r,t+1}$$

Also, $U^i_{r,1} = U_1\ \hat{p}^i_r$ (both known numbers, explained in Incoming Patient characteristics, Section 4.1). Then,$\underline{U}^i_{r,1} = U^i_{r,1}$ and we recursively build in the following manner. For *t=1*,

$$\underline{U}^i_{r,2} = \left(\sum_{i \in \mathcal{I}_w} \theta^{j,i}_{r,untreat} \, (1 - \alpha\beta) \underline{U}^i_{r,1} \right) + N^i_{r,2}$$

We repeat this step for all$t$. Next, to calculate$\overline{U}^i_{r,t}$, we run the QMPBB for $S_{r,t} = 0$ and use the $U^i_{r,t}$obtained from its solution to set$\overline{U}^i_{r,t} = U^i_{r,t}$.

References

Abadie, J., Carpentier, J. 1969. Generalization of the Wolfe reduced gradient method to the case of nonlinear constraints. *Optimization*, 37-47.

Anaya, H.D., Chan K., Karmarkar, U.S, Asch, S.M., Goetz, M.B. 2012. Budget Impact Analysis of HIV Testing in the VA Healthcare System.*Value in Health*, **15**, 1022-1028.

Anderson, R. M., May R.M., and Anderson B. 1992. *Infectious diseases of humans: dynamics and control*. Vol. 28. Oxford University Press, United Kingdom.

Bishai, D., Colchero, A., & Durack, D. T.2007 The cost effectiveness of antiretroviral treatment strategies in resource-limited settings. *AIDS*,*21*(10), 1333-1340.

Blount, S., Galambosi, A., & Yakowitz, S. 1997. Nonlinear and dynamic programming for epidemic intervention. *Applied Mathematics and Computation*, *86*(2), 123-136.

Brandeau, M. L., Zaric, G. S., & Richter, A. 2003. Resource allocation for control of infectious diseases in multiple independent populations: beyond cost-effectiveness analysis. *Journal of Health Economics*, *22*(4), 575-598.

Deo, S., Iravani, S., Jiang, T., Smilowitz, K., & Samuelson, S. Improving access to community-based chronic care through improved capacity allocation. Forthcoming, *Operations Research*.

Dolan, E., Fourer, R., Moré, J. J., Munson, T. S. 2002. The NEOS server for optimization: Version 4 and beyond. *Preprint ANL/MCS-TM-253, Mathematics and Computer Science Division, Argonne National Laboratory*.

Dolan, P., Shaw, R., Tsuchiya,A. and Williams, A. 2005. QALY maximization and people's preferences: a methodological review of the literature. *Health Economics,* 14.2 (2005): 197-208.

Drud, A. 1985. CONOPT: A GRG code for large sparse dynamic nonlinear optimization problems. *Mathematical Programming*, **31**(2), 153-191.

Freedberg, K. A., Scharfstein, J. A., Seage III, G. R., Losina, E., Weinstein, M. C., Craven, D. E., Paltiel, A. D. 1998. The cost-effectiveness of preventing AIDS-related opportunistic infections. *JAMA: the journal of the American Medical Association*, ***279***(2), 130-136.

Gandhi, N. R., Skanderson, M., Gordon, K. S., Concato, J., Justice, A. C. 2007. Delayed presentation for human immunodeficiency virus (HIV) care among veterans: a problem of access or screening?*Medical Care*, **45**(11), 1105.

Gidwani, R., Goetz, M. B., Kominski, G., Asch, S., Mattocks., K, Samet, J. H., Justice, A., Gandhi, N., and Needleman, J. 2009. A budget impact analysis of rapid human immunodeficiency virus screening in Veterans administration emergency departments. *The Journal of Emergency Medicine*.

Goetz, M. B., Bowman, C., Hoang, T., Anaya, H., Osborn, T., Gifford, A. L., Asch, S. M. 2008a. Implementing and evaluating a regional strategy to improve testing rates in VA patients at risk for HIV, utilizing the QUERI process as a guiding framework: QUERI Series. *Implement Sci*, *3*(1), 16.

Goetz, M. B., Smith, R., Teresa Osborn RN, M. S. N., Gifford, A. L., Asch, S. M. 2008b. A system-wide intervention to improve HIV testing in the Veterans Health Administration. *Journal of General Internal Medicine*, *23*(8), 1200-1207.

Goetz, M.B and Rimland, D.B. 2011. Effect of Expanded HIV Testing Programs in the Status of Newly Diagnosed HIV-Infected Patients in Two Veterans Health Administration Facilities. *JAIDSJournal of Acquired Immune Deficiency Syndromes,* 57(2), e23-e25.

Goetz, M. B., Hoang, T., Knapp, H., Henry, R.,Anaya, H., Chou, A.F.., Gifford, A. L., Asch, S. M. 2011. Exportability of an Intervention to Increase HIV testing in the Veterans Health Administration. *Journal of Quality and Patient Safety*, *37*(12), 553-559.

Goetz, M. B., Herschel Knapp PhD, M. S. S. W., Jane Burgess RN, M. S., MHP, M. D. F., Gifford, A. L., Asch, S. M. 2013. Central Implementation Strategies Outperform Local Ones in Improving HIV Testing in Veterans Healthcare Administration Facilities. *Journal of General Internal Medicine*, 1-7.

Gurobi Optimization, Inc. Gurobi Optimizer Reference Manual Version 3.0. Houston, Texas: Gurobi Optimization, April 2010.

Kahn, J. G., Brandeau, M. L., & Dunn-Mortimer, J. 1998. OR modeling and AIDS policy: from theory to practice. *Interfaces*, *28*(3), 3-22.

Kaplan, J. E., Benson, C., Holmes, K. K., Brooks, J. T., Pau, A., Masur H., 2009. "Guidelines for prevention and treatment of opportunistic infections in HIV-infected adults and adolescents."*MMWR Recomm Rep*58, no. RR-4: 1-207.

Kleinrock, L. 1975. Queuing Systems, Volume 1: Theory. John Wiley & Sons, Inc. Canada.

Kucukyazici, B., Verter, V., and Mayo, N. 2011. An analytical framework for designing community-based care for chronic diseases. *Production and Operations Management*, 20(3):474–488.

Long, E. F., Brandeau, M. L., & Owens, D. K. 2010. The cost-effectiveness and population outcomes of expanded HIV screening and antiretroviral treatment in the United States. *Annals of Internal Medicine*, 153(12), 778-789.

Mauskopf, J., Kitahata, M., Kauf, T., Richter, A., Tolson, J. 2005. HIV antiretroviral treatment: early versus later. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, **39**(5), 562-569.

Martin E.G., Paltiel A.D., Walensky R.P., Schackman B.R. 2010. Expanded HIV screening in the United States: what will it cost government discretionary and entitlement programs? A budget impact analysis. *Value Health*,**13**(8):893-902.

McCormick, G. P. 1976. Computability of global solutions to factorable non convex programs: Part I—Convex underestimating problems. *Mathematical Programming*, **10**(1), 147-175.

Nayak, S. U., Welch, M. L., Kan, V. L. 2012. Greater HIV testing after Veterans Health Administration policy change: the experience from a VA Medical Center in a high HIV prevalence area. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, *60*(2), 165-168.

Owens, D. K., Sundaram, V., Lazzeroni, L. C., Douglass, L. R., Sanders, G. D., Taylor, K., Holodniy, M. 2007. Prevalence of HIV infection among inpatients and outpatients in Department of Veterans Affairs health care systems: implications for screening programs for HIV. *Journal Information*, *97*(12).

Palella, Frank J.,Deloria-Knoll, M., Chmiel, J. S., Moorman, A. C., Wood, K. C., Greenberg, A., Holmberg, S., 2003 Survival benefit of initiating antiretroviral therapy in HIV-infected persons in different CD4+ cell strata. *Annals of Internal Medicine* 138, no. 8: 620-626.

Paltiel, A. D., Weinstein, M., C., Kimmel, A., D., Seage III, G., R., Losina, E., Zhang, H., Freedberg, K., A., Walensky, R. P., 2005. Expanded screening for HIV in the United States—an analysis of cost-effectiveness. *New England Journal of Medicine*, **352**(6): 586-595.

Petersen L. A., Urech T. H., Byrne M. M., Pietz K. 2007. Do financial incentives in a globally budgeted healthcare payment system produce changes in the way patients are categorized? A five-year study. *AJMCAmerican Journal of Managed Care*; **13**(9):513-22.

Rauner, M. S., & Brandeau, M. L. 2001. AIDS policy modeling for the 21st century: an overview of key issues. *Health Care Management Science*, *4*(3), 165-180.

Roberts, M. S., Nucifora, K. A., Braithwaite, R. S. 2010. Using mechanistic models to simulate comparative effectiveness trials of therapy and to estimate long-term outcomes in HIV care. *Medical Care*, *48*(6), S90-S95.

Schackman, B. R., Gebo, K. A., Walensky, R. P., Losina, E., Muccio, T., Sax, P. E., Weinstein, M. C.,Seage III, G. R., Moore, R. D., Freedberg, K. A.2006. The lifetime cost of current human immunodeficiency virus care in the United States. *Medical Care,* **44**(11) 990-997.

Shechter, S. M., Bailey, M. D., Schaefer, A. J., Roberts, M. S. 2008. The optimal time to initiate HIV therapy under ordered health states. *Operations Research*, *56*(1), 20-33.

Silva, A., Glick, N., R., Lyss, S., B., Hutchinson, A., B., Gift, T., L., Pealer, L., N., Broussard, D., Whitman, S. 2007. Implementing an HIV and sexually transmitted disease screening program in an emergency department. *Annals of Emergency Medicine,* **49**(5) :564-572.

Vishvanathan J., Grosssman, I.E. 1990. A combined Penalty Function and Outer Approximation Method for MINLP Optimization. *Computers and Chemical Engineering*, **14**, pp. 769–782

Wicaksono, D. S., Karimi, I. A. 2008. Piecewise MILP under-and over estimators for global optimization of bilinear programs. *AIChE Journal*, ***54***(4), 991-1008.

Zaric, G. S., Brandeau, M. L., & Barnett, P. G. 2000. Methadone maintenance and HIV prevention: a cost-effectiveness analysis. *Management Science, 46(8), 1013-103.*

Zaric, G. S., and Brandeau, M. L. 2001, Resource Allocation for Epidemic Control Over Short Time Horizons.*Mathematical Biosciences*, 171(1), 33-58.

Tables and Figures

Table 1 Health States

Health State Index ($i.j$)	CD4 Count Range (cells/mm^3) *without* Opportunistic Infections	Health State Index ($i.j$)	CD4 Count Range (cells/mm^3) *with* Opportunistic Infection
0	Uninfected	7	500+
1	500+	8	350-499
2	350-499	9	200-349
3	200-349	10	100-199
4	100-199	11	50-99
5	50-99	12	0-49
6	0-49	13	Death

Table 2: Notations

Indices	
$\tau \in [T] = \{1, 2, \ldots, T\}$	Number of years
$t \in \mathcal{M}_\tau = \{1 + 12(\tau - 1), \ldots, \tau\}$	Number of months
$k \in \mathcal{W} = \{phys, nurse, couns, lab\}$	Resource type
$l \in \mathcal{L} = \{P, L, S\}$	Location within health care system. P denotes primary care facility, L denotes laboratory and S denotes infectious diseases subspecialty.
$i, j \in \mathcal{I}_w = \{0,1, \ldots, 6\}$	Health states corresponding to patients without OI
$i, j \in \mathcal{I}_o = \{7, 8, \ldots, 13\}$	Health states corresponding to patients with OI
$i, j \in \mathcal{I}_w \cup \mathcal{I}_o = \mathcal{I} = \{0,1, \ldots, 13\}$	Health state of all patients
$\omega \in \mathcal{S} = \{treat, untreat\}$	Treatment status
$r \in \{1, 2\}$	Risk category
$X \in \mathcal{X} = \{U, W, E, M, D\}$	System state, U: Unscreened , W: Waiting for results, E: Waiting to be enrolled into monitoring or treatment, M: Monitoring, D: Treatment
Parameters (related to patient flow)	
$\hat{p}_r^i$	Fraction of patients in risk category r of health state i in the new patient population
α	Fraction of asymptomatic patients who visit health care facility
β	Fraction of patients who accept screening
$\theta_{r,\omega}^{i,j}$	Fraction of patients in risk category r and under treatment status ω moving from health state i to health state j in one month.
q^i	Quality of life score for patients in health state i
N_t	Number of new patients entering the system in period t
z^i	A binary parameter indicating whether patient of health state i is initiated under monitoring ($z^i = 0$) or treatment ($z^i = 1$)
Parameters (related to resource utilization)	
$y_{\mathrm{k,l}}$	Time required per patient of health care worker of type k at location l
$A_{\mathrm{k,l}}$	Total time available for HIV screening program of health care worker of type k at location l

w_k	Per period wages of health care worker of type k
CS^i	Cost of screening per patient
C_X^i	Cost per patient in system state X
$B(\tau)$	Total annual budget available for HIV related activities in year τ
State variables	
$U_{r,t}^i$	Number of unscreened patients of risk category r in health state i at the beginning of period t
$W_{r,t}^i$	Number of patients of risk category r in health state i waiting for their results at the beginning of period t
$R_{r,t}^i$	Number of patients of risk category r in health state i who receive their results in period t
$E_{r,t}^i$	Number of patients of risk category r in health state i waiting to be enrolled at the beginning of period t
$M_{r,t}^i$	Number of patients of risk category r in health state i who are under monitoring at the beginning of period t
$D_{r,t}^i$	Number of patients of risk category r in health state i who are under treatment at the beginning of period t
$ID_{r,t}^i$	Number of patients of risk category r in health state i who are initiated under treatment in period t
$IM_{r,t}^i$	Number of patients of risk category r in health state i who are initiated under monitoring in period t
$I_{r,t}^i$	Number of patients of risk category r who are initiated under care (monitoring and treatment) in period t
Decision Variables	
$S_{r,t}$	Fraction of asymptomatic patients of risk category r visiting a primary care facility in period $t$who are screened or offered the HIV test
$n_{k,l}$	Number of health care workers of type k to be staffed at location l

Table 3: QOL Weights

Health State (i)	QOL Weight (q^i)*	Health State (i)	QOL Weight (q^i)**
0	1	5	0.81
1	0.94	6	0.79
2	0.94	7-12	0.60
3	0.94	13	0
4	0.87		

Sources : * Mauskopf et al. (2005), ** Freedberg et al. (1998).

Table 4: % Gap of Heuristics and % Improvement from Current Practice

	Budget Level : Low		Budget Level : Medium		Budget Level : High	
	% Gap	% Improvement	% Gap	% Improvement	% Gap	% Improvement
FSSS	0	20.18	0.08	23.39	1.27	38.80
FSNS	0	20.21	0.2	24.13	1.33	40.15
VSSS	4.32	283.90	3.25	66.47	0.48	41.53
VSNS	7.05	305.30	5.15	69.69	3.9	42.94

Figure 1: Flow of patients through different parts of the health care system in the Greater Los Angeles

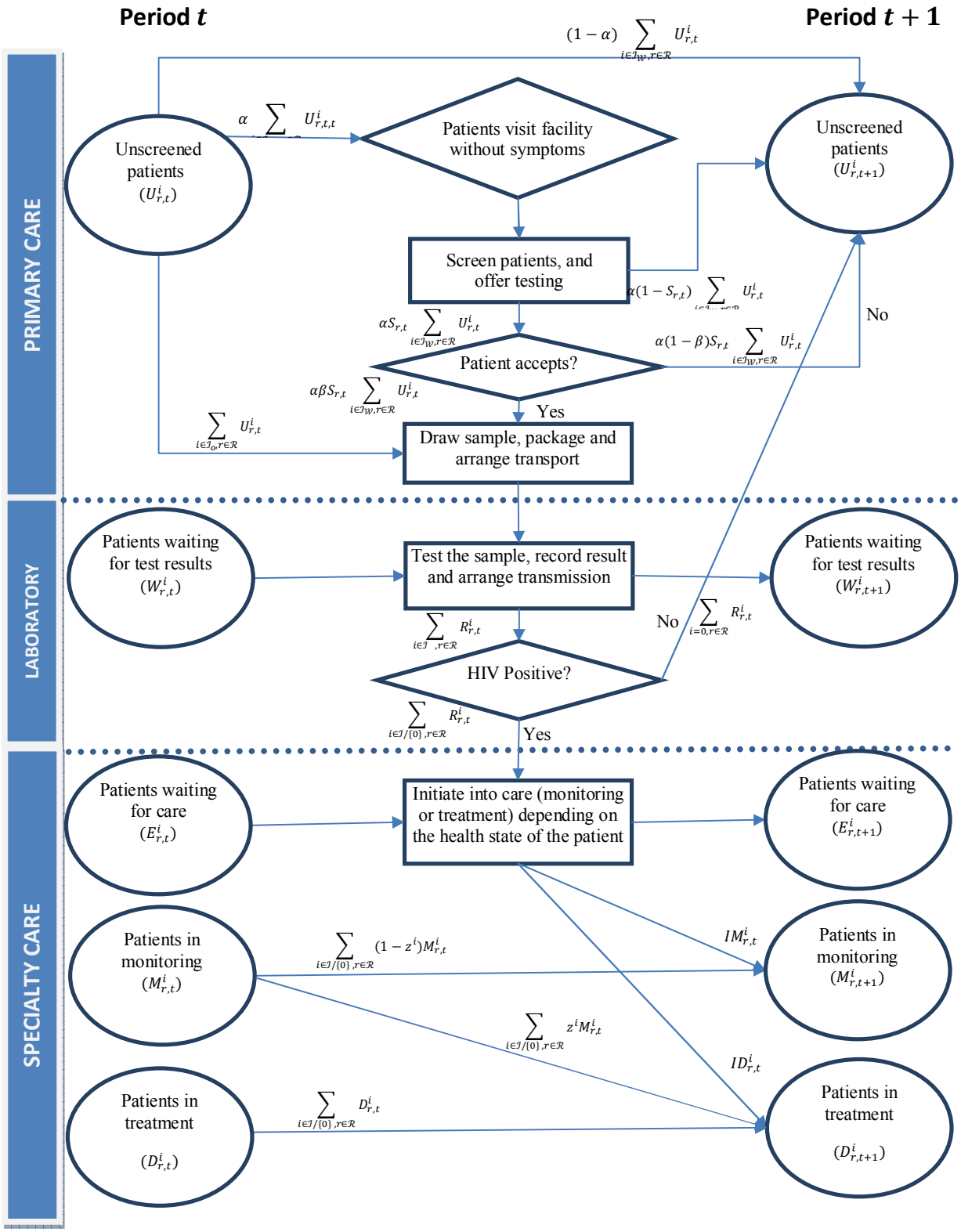


Figure 2: Screening rates over time for the non-stationary screening rate policies

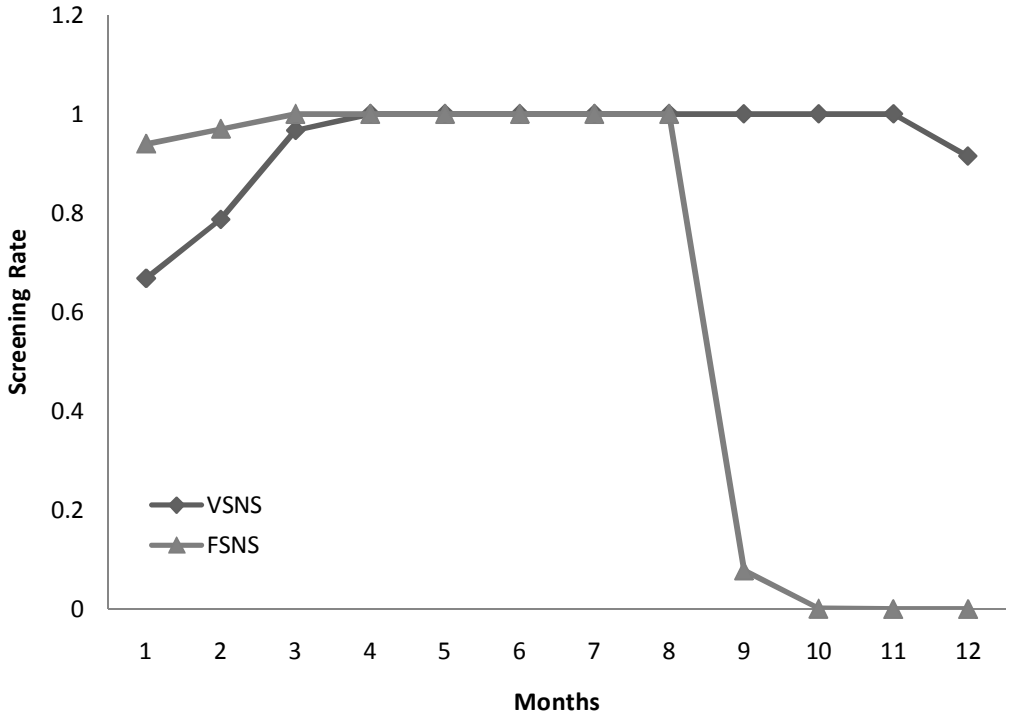
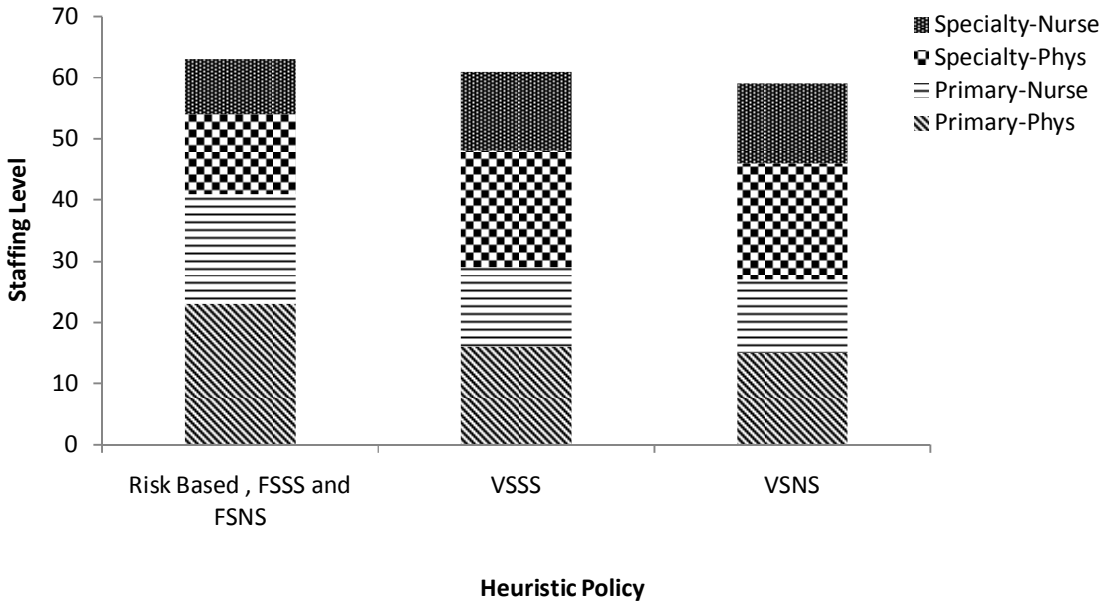


Figure 3: Staffing allocation across policies



ELECTRONIC COMPANION

Planning for HIV Screening, Testing and Care at the Veterans Health Administration

Sarang Deo
Indian School of Business, Gachibowli, Hyderabad, India 500032
(sarang_deo@isb.edu)
Kumar Rajaram, Sandeep Rath, Uday Karmarkar
UCLA Anderson School of Management, Los Angeles CA 90095
(kumar.rajaram@anderson.ucla.edu, sandeep.rath.2015@anderson.ucla.edu
uday.karmarkar@anderson.ucla.edu)
Matthew Goetz
Veteran's Health Administration, Greater Los Angeles Station
(matthew.goetz@va.gov)

1. Proof of Propositions

Proof of Proposition 1: We first use induction on t to show the following equation (16) holds. Let $K_{r,t}^i$ and $v_{r,t}^i$ be constants, then:

$$U_{r,t}^i + W_{r,t}^i + E_{r,t}^i + M_{r,t}^i + D_{r,t}^i = K_{r,t}^i + \sum_{i \in \mathcal{I}} \sum_{s \in \{1,2,\ldots,t\}} v_{r,s}^i D_{r,s}^i \; \forall \, i, r, t, \tag{16}$$

Observe that for $t = 1$, (16) is trivially true, since, $U_{r,1}^i$ is a constant and $W_{r,t}^i, E_{r,t}^i, M_{r,t}^i, D_{r,t}^i$ are all zero. Next, assume that (16) holds for t. We show that (16) then holds for $t+1$. From (1), (2), (3), (6) and (7) for $t+1$ we get,

$$\begin{aligned} U_{r,t+1}^i + W_{r,t+1}^i + E_{r,t+1}^i + M_{r,t+1}^i + D_{r,t+1}^i \\ &= \sum_{j \in \mathcal{I}} \left(U_{r,t}^j + W_{r,t}^j + E_{r,t}^j + M_{r,t}^j + D_{r,t}^j\right) \theta_{r,untreat}^{ji} + \sum_{j \in \mathcal{I}} D_{r,t}^j \left[\theta_{r,treat}^{ji} - \theta_{r,untreat}^{ji}\right] \\ &\quad + N_{r,t+1}^i \\ &= \sum_{j \in \mathcal{I}} \left(K_{r,t}^j + \sum_{h \in \mathcal{I}, s \in \{1,2,\ldots,t\}} v_{r,s}^{hj} D_{r,s}^h \right) \theta_{r,untreat}^{ji} + \sum_{j \in \mathcal{I}} D_{r,t}^j \left[\theta_{r,treat}^{ji} - \theta_{r,untreat}^{ji}\right] + N_{r,t+1}^i \\ &= K_{r,t+1}^i + \sum_{j \in \mathcal{I}, s \in \{1,2,\ldots,t+1\}} v_{r,s}^{ji} D_{r,s}^j \end{aligned}$$

Where, $K^i_{r,t+1} = \sum_{j \in \mathcal{I}} \; K^j_{r,t}\,\theta^{ji}_{r,untreat} + N^i_{r,t+1}, \nu^{ji}_{r,t+1} = \left[\theta^{ji}_{r,treat} - \theta^{ji}_{r,untreat}\right]$ and $\nu^{ji}_{r,t} = \sum_{h\in\mathcal{I},j\in\mathcal{I},s\in\{1,2,\ldots,t\}} \nu^{hj}_{r,s}\theta^{ji}_{r,untreat}$

This shows that if (16) holds for t it also holds for $t+1$. Therefore, by induction (16) is true. We next substitute (16) in the objective function of the QMPPB to get:

$\sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i \left(K^i_{r,t} + \sum_{j\in\mathcal{I},} \; \nu^{ji}_{r,s} D^j_{r,s} \right)$. Simplifying, we have:

$$K_0 + \sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau} \; \pi^i_{r,t} D^i_{r,t},$$

Where, $K_0 = \sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i K^i_{r,t}$. and $\pi^i_{r,t} = \sum_{j\in\mathcal{I},s\in\{1,2,\ldots,t\}} \; q^i \nu^{ji}_{r,s}$. ■

Proof of Proposition 2 Consider inequality (9) of the QMPP:

$$\sum_{j\in\mathcal{I}_w,r\in\mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R}} U^i_{r,t} \le min_k\left\{n_{k,P}A_{k,P} \; / y_{k,P}\right\} + \frac{1}{\tau_P}\ln(1-\alpha_P)$$

This can be written as:

$$\sum_{i\in\mathcal{I}_o,r\in\mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i\in\mathcal{I}_W/\{13\},r\in\mathcal{R}} U^i_{r,t} \le n_{k,P}A_{k,P} \; / y_{k,P} + \frac{1}{\tau_P}\ln(1-\alpha_P) \qquad \forall k$$

Replacing $U^i_{r,t}$ with its lower bound $\underline{U}^i_{r,t}$, and rearranging terms, we get:

$$\left(y_{k,P}/A_{k,P}\right)\left(\sum_{j\in\mathcal{I}_w,r\in\mathcal{R}} \alpha\beta S_{r,t}\underline{U}^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R}} \underline{U}^i_{r,t} \right) - \frac{y_{k,P}}{A_{k,P}\;\tau_P}\ln(1-\alpha_P) \le n_{k,P}\forall k$$

Multiplying each term by w_k and summing across k constraints:

$$\sum_{k\in W}\left\{\left(w_k\, y_{k,P}/A_{k,P}\right)\left(\sum_{j\in\mathcal{I}_w,r\in\mathcal{R}}\alpha\beta S_{r,t}\underline{U}^i_{r,t}+\sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R}}\underline{U}^i_{r,t}\right)-\frac{w_k y_{k,P}}{A_{k,P}\;\tau_P}\ln(1-\alpha_P)\right\}$$

$$\leq\sum_{k\in W}w_k n_{k,P} \qquad (1E)$$

Similarly using inequality (11) of the QMPP, we get:

$$\sum_{k\in W}\left\{\left(w_k y_{k,S}/A_{k,S}\right)\sum_{i\in\mathcal{I}/\{0\},r\in\mathcal{R}}D^{\mathrm{i}}_{\mathrm{r,t}}\varphi^i_D-\frac{w_k y_{k,S}}{A_{k,S}\;\tau_S}\ln(1-\alpha_S)\right\}$$

$$\leq\sum_{k\in W}w_k n_{k,S} \qquad (2E)$$

Consider inequality (8) of the QMPP:

$$\sum_{i\in\mathcal{I}_w,r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^i\alpha\beta S_{r,t}U^i_{r,t}+\sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^iU^i_{r,t}+\sum_{i\in\mathcal{I}\ /\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau,X\in\mathcal{X}}C^i_X X^i_{r,t}$$

$$+\sum_{l\in\mathcal{L},k\in W,t\in\mathcal{M}_\tau}n_{k,l}w_k\leq B(\tau)$$

Replacing $U^i_{r,t}$ with its lower bound $\underline{U}^i_{r,t}$, and since the other terms are positive, we get:

$$\sum_{i\in\mathcal{I}_w,r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^i\alpha\beta S_{r,t}\underline{U}^i_{r,t}+\sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^i\underline{U}^i_{r,t}+\sum_{i\in\mathcal{I}\ /\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}C^i_U\underline{U}^i_{r,t}$$

$$+\sum_{i\in\mathcal{I}\ /\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}C^i_D D^i_{r,t}+\sum_{l\in\mathcal{L},k\in W,t\in\mathcal{M}_\tau}n_{k,l}w_k\leq B(\tau)$$

Substituting (1E) and (2E) in the above inequality:

$$\sum_{i\in\mathcal{I}_w,r\in\mathcal{R},t\in\mathcal{M}_\tau} CS^i\alpha\beta S_{r,t}\underline{U}^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau} CS^i\underline{U}^i_{r,t} + \sum_{i\in\mathcal{I}\ /\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau} C^i_U\underline{U}^i_{r,t}$$

$$+ \sum_{i\in\mathcal{I}\ /\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau} C^i_D D^i_{r,t}$$

$$+ \sum_{k\in W,t\in\mathcal{M}_\tau}\left\{\left(w_k\, y_{k,P}/A_{k,P}\ \right)\left(\sum_{i\in\mathcal{I}_w,r\in\mathcal{R}}\alpha\beta S_{r,t}\underline{U}^i_{r,t} + \sum_{i\in\mathcal{I}_o/\{13\},r\in\mathcal{R}}\underline{U}^i_{r,t}\right) - \frac{w_k y_{k,P}}{A_{k,P}\ \tau_P}\ln(1\right.$$

$$\left. - \alpha_P)\right\} + \sum_{k\in W,t\in\mathcal{M}_\tau}\left\{\left(w_k y_{k,S}/A_{k,S}\ \right)\sum_{i\in\mathcal{I}_w,r\in\mathcal{R}} D^i_{r,t}\varphi^i_D - \frac{w_k y_{k,S}}{A_{k,S}\ \tau_S}\ln(1-\alpha_S)\right\} \le B(\tau)$$

This simplifies to

$$\sum_{r\in\mathcal{R},t\in\mathcal{M}_\tau}\sigma_{r,t}S_{r,t} \le B(\tau) - K_\tau - \sum_{i\in\mathcal{I}\ /\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}\rho^i D^i_{r,t} \quad \forall\tau \qquad (3E)$$

This is the first inequality in the proposition with the associated definitions of K_τ, ρ^i and $\sigma_{r,t}$. Note that the total number of patients treated in each risk category, has to be less than the total number of patients screened and the total number of unscreened patients who get infected with OI. Thus:

$$\sum_{i\in\mathcal{I},t\in\mathcal{M}_\tau} D^i_{r,t} \le \sum_{j\in\mathcal{I}_w,t\in\mathcal{M}_\tau}\alpha\beta S_{r,t}\overline{U}^i_{r,t} + \sum_{j\in\mathcal{I}_o,t\in\mathcal{M}_\tau}\overline{U}^i_{r,t} \quad \forall r$$

The above inequality can be rearranged to get the second inequality in the proposition.

For stationary screening, setting $S_{r,t} = S_r\ \forall t$ and since $\rho^i D^i_{r,t} \ge 0$ and $\sigma_{r,t}S_{r,t} \ge 0$, from (3E) we get $S_r \le \frac{B(\tau)-K_\tau}{\sum_{t\in\mathcal{M}_\tau}\sigma_{r,t}}$■

2. Search Algorithm for Stationary Screening.

Start: ΔS, $i \leftarrow 0$, $j \leftarrow 0$, $S_{hi} \leftarrow 0$, $S_{lo} \leftarrow 0$, max$\leftarrow$ 0, N$\leftarrow$ [1/ΔS], $\bar{S}_{lo} = min_{\tau}\left\{\frac{B(\tau)-K_{\tau}}{\sum_{t\in\mathcal{M}_{\tau}}\sigma_{lo,t}}\right\}$, $\bar{S}_{hi} = min_{\tau}\left\{\frac{B(\tau)-K_{\tau}}{\sum_{t\in\mathcal{M}_{\tau}}\sigma_{hi,t}}\right\}$

While i<N+1 and $S_{lo} \leq \bar{S}_{lo}$

Do,

$S_{lo} \leftarrow S_{lo}+i\Delta S$

$j \leftarrow 0$

While j<N+1 and $S_{hi} \leq \bar{S}_{hi}$

Do,

$S_{hi} \leftarrow S_{hi}$+jΔS

Evaluate QMPPB(S_{hi}, S_{lo})

If QMPPB is infeasible

then,

end do

If max< QMPPB(S_{hi}, S_{lo}),

then

max$\leftarrow$ QMFS(S_{hi}, S_{lo})

$Sopt_{hi} \leftarrow S_{hi}$

$Sopt_{lo} \leftarrow S_{lo}$

else

$j \leftarrow j$+1

end do

end do

return max, $Sopt_{hi}, Sopt_{lo}$

3. Budget Imputation Algorithm

Start:

set $S_{lo,t} \leftarrow 0, S_{hi,t} \leftarrow 1$, $\Delta B \leftarrow \$0.5mn$, *count*$\leftarrow$*0*, *B*$\leftarrow$*0, obj,* $\underline{B} \leftarrow$0, maxQ$\leftarrow$ 0, $\bar{B} \leftarrow 0$, exit$\leftarrow$ 0

/*to calculate $\underline{B}$ */

while (exit=0)

do,

evaluate QMPPB($S_{r,t}, B$)

if QMPPB($S_{r,t}, B$) is feasible

then,

$\underline{B} \leftarrow$QMPPB($S_{r,t}, B$)

exit$\leftarrow$ 1

else,

count$\leftarrow$ count+1

$B \leftarrow B + \Delta B * count$

end if
end do

/*to calculate $\bar{B}$*/

exit← 0 /*re-initialize exit flag*/

$B \leftarrow \underline{B}$
while (exit=0)
do,
evaluate QMPPB($S_{r,t}, B$)
if maxQ>QMPPB($S_{r,t}, B$)
then,

$$\bar{B} \leftarrow \text{QMPPB}(S_{r,t}, B)$$

exit← 0
else,
count← count+1

$$B \leftarrow B + \Delta B * count$$

end if
end do

4. Procedure for choosing the number of partitions in the upper bound calculation

In determining the upper bound for the QMPP we need to choose parameter m, the number of partitions on $U_{r,t}^{i}$. Note that as m increases, the value of the upper bound decreases (or becomes tighter) but its computation time becomes larger. Our procedure chooses m by comparing this reduction of the bound value with its increase in computation time. To initialize this procedure, we start with $m = 1$ and record the value of the upper bound along with the time GUROBI takes to compute the bound. Next, we increment m by 1 and calculate the % reduction of the value of the bound and the % increase in computation time from the previous value of m. We then calculate the efficiency ratio defined as (% reduction in bound value)/(% increase in computation time) and choose m corresponding to the highest ratio. We applied this procedure to our data for $m = 1\ to\ 7$ as GUROBI was unable to solve upper bounds for $m > 7$. We found the best choice was at $m = 5$.

5. Estimation of system state costs C_X^i

C_X^i is composed of the following:

1) In Patient costs (CI^i): The average in-patient costs, (CI^i) per patient per month was collected from VHA data. This cost is incurred on all the patients at each system state. Thus the in-patient cost is:

$$CI^i\left(\alpha U_{r,t}^i + W_{r,t}^i + E_{r,t}^i + M_{r,t}^i + D_{r,t}^i\right) \qquad (1E)$$

2) Monitoring costs (CM^i): The monthly per-patient monitoring costs CM^i, is incurred on patients under monitoring $M^i_{r,t}$, as well as treatment $D^i_{r,t}$. This is the cost of one CD4 cell count and one HIV-1 RNA quantitation, per quarter. Anaya et al. (2012) provide the cost of CD4 cell count and RNA quantitation. The monitoring is:

$$CM^i\left(M^i_{r,t} + D^i_{r,t}\right) \qquad (2E)$$

3) Treatment costs (CT^i): The treatment cost per patient CT^i is the cost of pharmacy for patients undergoing treatment under HAART. The treatment cost is :

$$CT^i D^i_{r,t} \qquad (3E)$$

4) Outpatient overhead costs (Coh^i_X): The per patient overhead costs, Coh^i_X , was not directly available. Only the per-patient outpatient cost CO^i, was available from VA. This cost however, was inclusive of monitoring test costs and labor costs, which have already been accounted in the monitoring costs described above and in wages. Thus, in order to calculate outpatient overhead costs, we need to subtract the monitoring costs and the labor cost is:

$$Coh^i_{X,M} = CO^i - CM^i - L^i_X$$

Here, L^i_X is the out-patient labor utilization cost per patient at system state X. Let $y_{X,k}$ denote the labor time of staff of type k, required per patient visit at system state X. Further let w_k denote the wage per time of staff type $k$and the $\varphi^i_X$the frequency of visits. These are them used to calculate the labor cost incurred per patient per month as

$$L^i_X = \varphi^i_M \sum_{k\epsilon W}\left(y_{k,X} w_k\right)$$

Since outpatient overhead cost is incurred on all patients in the system, the total outpatient overhead cost for year τ would be given by:

$$\sum_{i\in\mathcal{I}/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}\left[\left(Coh^i_U \alpha U^k_{r,t} + Coh^i_W W^k_{r,t} + Coh^i_E E^k_{r,t} + Coh^i_M M^k_{r,t} + Coh^i_D D^k_{r,t}\right)\right] \qquad (4E)$$

Summing equations (1E) through (3E) over all time periods, risk categories and health states and adding equation (4E), we get:

$$\sum_{i\in\mathcal{I}/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}\Big[\Big(\left(Coh^i_U \alpha + CI^i \alpha\right)U^i_{r,t} + (Coh^i_W + CI^i)W^i_{r,t} + (Coh^i_E + CI^i)E^i_{r,t}$$

$$+ (Coh^i_M + CM^i + CI^i)M^i_{r,t} + (Coh^i_D + CI^i + CM^i + CT^i)D^i_{r,t}\Big)\Big]$$

Collecting the terms in order to simplify the notation the total costs can be written as,

$$\sum_{i\in\mathcal{I}/\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau,X\in\mathcal{X}} [C_X^i X_{r,t}^i]$$

where,

$$C_U^i = \alpha(CI^i + Coh_U^i)$$
$$C_W^i = Coh_W^i + CI^i$$
$$C_E^i = Coh_E^i + CI^i$$
$$C_M^i = Coh_M^i + CI^i + CM^i$$
$$C_D^i = Coh_D^i + CI^i + CM^i + CT^i$$

For brevity, we report $C_X^i\ \forall i, X$ in Table 1A. Detailed breakdown are available upon request from the authors.

6. Computation of Transition Rates

As discussed in the paper, there are four processes which govern the transition from one health state to another: 1) HIV infection, 2) HIV infection progression (treated and untreated), 3) Opportunistic infection (OI), and 4) OI recovery.

The first process is the HIV infection process which governs the transition from health state 0 (uninfected) to health state 1 (>500cells/mm^3). The monthly rate of transition under the HIV Infection process is denoted by $\theta_{r,untreat}^{0,1}$, where $\theta_{r,untreat}^{0,1} = incid_r/12$, where $incid_r$ is the annual incidence rate of risk category r.We used the estimates provided by Paltiel et al. (2005) for the incidence rates($incid_r$) This is shown in Table 3A in the Electronic Companion.

The HIV progression process governs progression from one infected state to a higher infected state. The transition rate of this process varies depending on whether the patient is undergoing Highly Active Anti-Retroviral Treatment (HAART) or not. This transition rate from infected stage i to infected stage j for risk category r is given by $\theta_{r,treat}^{ij}$ and $\theta_{r,untreat}^{ij}$ for patients under HAART and not under treatment respectively. Mauskopf et al. (2005), calculate $p_{6-month}^{i,j}$, the six month transition probabilities from one health state to another *without* treatment. These 6 month transition probabilities are used to calculate monthly rate as $\theta_{r,untreat}^{i,j} = 1 - (1 - p_{6-month}^{i,j})^{1/6}$. These transition rates are tabulated in Table 8A.

Mauskopf et al. (2005) also provide relative risk of transition ($relrisk_{TR}^{i,j}$) between states in different treatment regimens (TRs), namely, First-line, Second-line, Salvage, and Optimized Background therapies, (Table 9A). This relative risk is used to calculate the transition rates under each treatment regimen. The transition rate under treatment regimen TR is given as $\theta_{r,TR}^{i,j} = \theta_{r,untreat}^{i,j}\left(1 - relrisk_{TR}\right)$. The overall transition rate under treatment is given by average of the transition rates under different treatment regimens or:

$$\theta_{r,treat}^{i,j} = \left(\theta_{r,first-line}^{i,j} + \theta_{r,second-line}^{i,j} + \theta_{r,salvage}^{i,j} + \theta_{r,optimized}^{i,j}\right)/4$$

The third process is the OI process that relates to patients infected with HIV who are susceptible to such infections. The rate with which they can be infected with these infections depends on the nature of the opportunistic infection and the current CD4 state of the patient. This transition rate is given by $\theta_{r,treat}^{i,i+6}$ and $\theta_{r,untreat}^{i,i+6}$ where $i \in \mathcal{I}_{\mathcal{W}}$. Paltiel et al. (2005) provide the monthly risk of being infected with OI by CD4 stratum and shown in Table 10A. For each CD4 category, we sum across the different OI to calculate the average risk of infection of OI. To illustrate, if we want to calculate $\theta_{r,\omega}^{2,8}$, we note from Table 1 that for $i = 2$and $j = 8$ correspond to a CD4 count between 350-499. We then go to this column in Table 10A and sum the appropriate column to get $2.27 \times 10^{-3} = \theta_{r,\omega}^{2,8}$.

Finally, the OI recovery process governs the recovery from such infection. The transition rates here are given by $\theta_{r,treat}^{i+6,i}$, where$i \in \mathcal{I}_{\mathcal{W}}$. Kaplan et al. (2009) provide typical time required for recovery from each OI as listed in Table 11A. As shown in this table, the typical recovery times are converted to a weighted average recovery time using the relative risk of incurring that OI. This weighted average monthly recovery time is converted to the fraction or rate of patients recovering every month by $1 - e^{-1.06} = 0.654$. Thus the transition rate from any OI infected state to OI uninfected state of the same CD4 bracket $\theta_{r,treat}^{i+6,i}$, $i \in \mathcal{I}_{\mathcal{W}}$is 0.654. Finally, due to the nature of HIV, $\theta_{r,untreat}^{i+6,i} = 0$, $i \in \mathcal{I}_{\mathcal{W}}$.

For transitions that require two processes to occur simultaneously such as transition between health states *and* transition to an OI status, we assume independence. Thus, the rates of the two processes occurring simultaneously are the product of the rates of the individual processes.

Finally, there are a total of four transition rate matrices corresponding to the two risk categories (i.e., high and low) and two treatment categories (treated vs. untreated). These transition rates are provided in Tables 12A through 15A.

7. Estimation of Quality of Life Utilities

The Quality of Life (QOL) utilities are drawn from two sources, Mauskopf et al. (2005) and Freedberg et al. (1998). Specifically, Mauskopf et al. provides 5 CD4 ranges, ≥500 cells/μL, 350-499 cells/μL, 200-349 cells/μL, 100-199 cells/μL and 0-100 cells/μL and death. We further divide the range 0-100 cells/μL into two, 50-99 cells/μL and 0-49 cells/μL because the treatment and system costs for these two CD4 ranges were different (Schackman et al., 2006). These health states are numbered 1 through 6 and death. The QOL utilities for health states 1-4 was from Table 2 in Mauskopf et al (2005). The QOL utilities for health states 5 and 6 were from Table 2 in Freedberg et al. By definition, the no infection state 0 has a QOL utility 1 and the death state 13 has a QOL utility 0.

Based on discussions with the physicians at the GLA station, we also incorporated health states with opportunistic infections by adding health states 7 through 12. As shown in Table 1, each of these states correspond to the same CD4 counts as in states 1 through 6 respectively, but have opportunistic infections. For example health state 7 (i.e., CD4 ≥500 cells/μL) corresponds to the CD4 count of health state 1, health state 8 with health state 2, and so on. The QOL utility for health states 7-12 were calculated from Table 2 in Freedberg et al. Here, we considered the health related quality adjustment scores for the opportunistic infections by listed pathogen types (such as Pneumocystis Carini, through other AIDS diagnoses). Ideally, one would have had to introduce additional sub health states for each opportunistic infection within a CD4 count range. However, the physicians felt that it would be impractical to do since patients typically had more than one opportunistic infection, it was often not easy to diagnose the pathogens and decide which one was most dominant. Further, the range of the scores across these opportunistic infections was relatively narrow (i.e., 0.56 to 0.65). Therefore, it was considered reasonable to calculate the quality utility for health states 7 through 12 by averaging the quality scores across these opportunistic infections.

8. Model Extension to Longer Time Horizons

The model can be easily extended to longer time horizons with the appropriate choice of T, where $\tau = \{1, 2, .., T\}$. To illustrate, we consider a five year and a ten year horizon. For the five year horizon we set T to 5, while in the 10 year horizon, we set T to 10. In both these cases, we use the upper bound developed in Section 3.4 and the heuristics developed in Section 3.5 of the paper. The percentage gaps and improvements from the risk based screening policy for the five year and ten year horizon are described in Tables 16A and 17A respectively. Note that these are very comparable to the analysis of the two year horizon as was reported in Table 4 of the paper.

To demonstrate the robustness and stability of the two year decision given a longer planning horizon, we used the solution of the two year problem in the five and ten year horizon across the different policies and budget levels. The reduction in the objective from its original value for the five year problem ranged from 2% to 5% averaging around 3%. Similarly, the reduction in the objective from its original value for the ten year problem ranged from 3% to 7% averaging around 5%. These results show that the two year solution is stable and robust. In fact, these reductions would be even lower if the model parameters are updated every year with the latest estimates as it would bedone in practice.

9. Impact of Early Screening on Budgets and QALYs gained.

Early screening could provide societal benefits by reducing transmission and ultimately prevalence rates. This is because when HIV infected individuals know their status via early screening, they are less likely to participate in unsafe sex and share syringes if they use intravenous drugs. However, it is not possible to *analytically* estimate this reduction as it depends on individual behavior (i.e., whether one would take adequate precautions after being diagnosed) and if the people affected by this individual are a part of the VHA system. Thus, to understand the benefits of early screening via reduced transmission to the general population, we systematically reduced prevalence rates by a fixed percentage in future periods. This reduction in prevalence rates affects parameter $\hat{p}_r^i$, the proportion of patients in each risk and CD4 category (Paltiel et al. 2005; Gandhi et al. 2007) and $N_{r,t}^i$, the number of new patients in each risk category and health state in each period who enter the station.We then used the risk based screening policy to calculate: 1) The change in budget to achieve the level of QALY's gained at the initial prevalence rate and 2) The change in QALYs gained if the budget levels are at the same level. These are summarized in Figure1A.This figure shows that even small reduction in prevalence rate can significantly reduce the budget required or increase the QALYs gained. We repeated this analysis for the policies described in Section 3.5 and obtained similar results. Thus, thisanalysis provides a model based justification for developing early screening programs.

Tables

Table 1A: System State Cost in $/per patient-month

	C_X^i				
Health State(i)	C_U^i	C_W^i	C_E^i	C_M^i	C_D^i
0	0.00	0.00	0.00	55.00	55.00
1	26.45	52.60	60.86	104.48	312.59

2	25.44	50.50	60.86	101.59	308.47
3	59.34	118.27	129.93	168.86	550.88
4	74.99	149.58	160.98	200.28	553.43
5	75.06	149.71	160.98	200.46	550.83
6	119.73	239.02	251.60	289.27	840.93
7-12	0.00	0.00	0.00	55.00	1820.70
13	0.00	0.00	0.00	0.00	0.00

Table 2A: Wages

Resource	Wage ($/month)
Physician	15,000
Nurse	11,000
Laboratory Assistant	7550
Counselor	6500

Table3A: Incidence and Prevalence Rate

Risk Category	$incid_r$	$prev_r$
1 (high risk)	0.012	0.03
2 (low risk)	0.0001	0.001

Table 4A: Incoming proportion by CD4 count

i	$\hat{p}_1^i$	$\hat{p}_2^i$
0	9.70E-01	9.99E-01
1	4.05E-03	1.35E-04
2	4.05E-03	1.35E-04
3	6.60E-03	2.20E-04
4	4.05E-03	1.35E-04
5	4.05E-03	1.35E-04
6	7.20E-03	2.40E-04
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0

Table 5A: Time Required by resource type and location y(k,l) in minutes/patient-visit

	Physician	Nurse	Lab Technician	Counselor
P	7	7	0	0
L	0	0	25	0

S	10	10	0	8

Table 6A: Time Available per resource per month (in minutes/month) A(k,l)

	P	L	S
phys	1600	0	2400
nurse	2000	0	3600
lab	0	6400	0
couns	0	0	3200

Table 7A: Out-patient visit frequency

i	Monitoring (φ_M^i)	Treatment (φ_D^i).
0	0.63	0.53
1	0.63	0.53
2	0.79	0.72
3	0.89	0.79
4	0.87	0.88
5	0.86	1
6	0.96	1
7	0.00	2.51
8	2.51	2.51
9	2.51	2.51
10	2.51	2.51
11	2.51	2.51
12	2.51	2.51
13	0	0

Table 8A: Transition Rates (Mauskopf et al. 2005)

State i to state j	Initial State to Final State	6 month rate	Monthly rate
$\theta_r^{1,2}$	500+ to 350-499	0.37	0.07294117
$\theta_r^{2,3}$	350-499 to 200-349	0.37	0.07294117
$\theta_r^{3,4}$	200-349 to 100-199	0.37	0.07294117
$\theta_r^{4,5}$	100-199 to 50-99	0.51	0.11134859
$\theta_r^{5,6}$	50-99 to <50	0.51	0.11134859
$\theta_r^{6,13}$	<50 to death	0.51	0.11134859

Table 9A: Relative Risk of Transition between States (Mauskopf et al. 2005)

	CD4+ gain	VL decrease	Relative risk of transition between states ($relrisk_{TR}$)
First-line	79	21.42	27%
Second-line	73	21.49	26.53%
Salvage therapy	76	21.697	22.80%
Optimized	32	20.763	51.95%

Table 10A: Transition Probability for OI (Paltiel et al. 2005)

	0 - 49/mm3	50- - 99/mm3	100 - 199/mm3	200 - 299/mm3	300 - 499/mm3	≥ 500/mm3
PCP	3.70E-02	3.10E-02	9.60E-03	3.73E-03	8.50E-04	4.10E-04
MAC	1.22E-02	3.75E-03	1.01E-03	2.20E-04	5.50E-05	5.90E-05
Toxoplasmosis	2.70E-03	1.40E-03	6.70E-04	4.20E-04	9.20E-05	2.90E-05
Cytomegalovirus	1.86E-02	5.23E-03	2.14E-03	5.80E-04	1.29E-04	5.90E-05
Fungal infection	1.12E-02	5.91E-03	1.35E-03	2.90E-04	2.76E-04	8.80E-05
Other	3.94E-02	2.46E-02	7.16E-03	2.24E-03	8.70E-04	4.70E-04
Total	**1.21E-01**	**7.19E-02**	**2.19E-02**	**7.48E-03**	**2.27E-03**	**1.12E-03**

Table 11A: Recovery Rates from OI

Infection	Days of Recovery	Monthly Rate (MR)	Weight (Wt)	Relative Risk (MR X Wt)	
PCP	21	1.42	0.082	0.1179	
MAC	14-28	1.42	0.0173	0.0247	
Toxoplasmosis	42	0.71	0.0053	0.0038	
Cytomegalovirus	21-28	1.22	0.027	0.0327	
Fungal Infection	70	0.42	0.019	0.0082	
Others	163	0.074	0.0274	0.002	
Weighted Average					1.06

Table 12A: Transition Rates High Risk , untreated $\theta^{i,j}_{1,untreat}$

i/j	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.999	0.001												
1		0.926	0.073					0.001	0.000					
2			0.925	0.073					0.002	0.000				
3				0.920	0.072					0.007	0.001			
4					0.869	0.109					0.019	0.002		
5						0.825	0.103					0.064	0.008	

6							0.781						0.108	0.111
7		0.606	0.048					0.321	0.025					
8			0.606	0.048					0.321	0.025				
9				0.606	0.048					0.321	0.025			
10					0.581	0.073					0.307	0.039		
11						0.581	0.073					0.307	0.039	
12							0.581						0.307	0.111
13														1.000

Table 13A: Transition Rates High Risk, Treated $\theta^{i,j}_{1,treat}$

i/j	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.999	0.001												
1		0.964	0.035					0.001	0.000					
2			0.963	0.035					0.002	0.000				
3				0.958	0.035					0.007	0.000			
4					0.925	0.053					0.021	0.001		
5						0.878	0.050					0.068	0.004	
6							0.831						0.115	0.054
7		0.631	0.023					0.334	0.012					
8			0.631	0.023					0.334	0.012				
9				0.631	0.023					0.334	0.012			
10					0.619	0.035					0.327	0.019		
11						0.619	0.035					0.327	0.019	
12							0.619						0.327	0.054
13														1.000

Table 14A: Transition Rates Low Risk, untreated $\theta^{i,j}_{2,untreat}$

i/j	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	1.000	0.000												
1		0.926	0.072					0.001	0.000					
2			0.925	0.072					0.002	0.000				
3				0.920	0.072					0.006	0.000			
4					0.869	0.108					0.019	0.002		
5						0.824	0.103					0.063	0.008	
6							0.781						0.107	0.111
7		0.60	0.047					0.320	0.025					

8			0.606	0.047					0.320	0.025				
9				0.606	0.047					0.320	0.025			
10					0.581	0.072					0.307	0.038		
11						0.581	0.072					0.307	0.038	
12							0.581						0.307	0.111
13														1.000 0

Table 15A: Transition Rates Low Risk, Treated $\theta_{2,treat}^{i,j}$

i/j	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	1.000	0.000												
1		0.964	0.035					0.001	0.000					
2			0.963	0.035					0.002	0.000				
3				0.958	0.035					0.007	0.000			
4					0.925	0.053					0.021	0.001		
5						0.878	0.050					0.068	0.004	
6							0.831						0.115	0.054
7		0.631	0.023					0.334	0.012					
8			0.631	0.023					0.334	0.012				
9				0.631	0.023					0.334	0.012			
10					0.619	0.035					0.327	0.019		
11						0.619	0.035					0.327	0.019	
12							0.619						0.327	0.054
13														1.000

Table 16A: % Gap of Heuristics and % Improvement from Current Practice for 5 year Horizon

	Budget Level : Low		Budget Level : Medium		Budget Level : High	
	% Gap	% Improvement	% Gap	% Improvement	% Gap	% Improvement
FSSS	0.17	10.25	0.05	18.26	4.56	27.45
FSNS	0.87	17.37	0.59	19.33	4.48	38.22
VSSS	6.77	195.25	5.64	57.25	2.56	30.63
VSNS	8.18	278.28	7.45	56.36	5.34	38.77

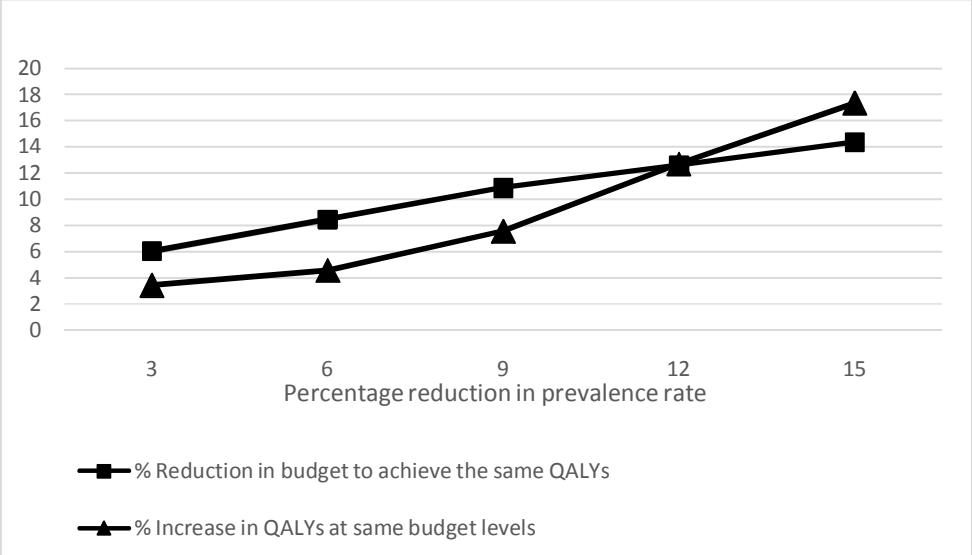
Table 17A: % Gap of Heuristics and % Improvement from Current Practice for 10 year Horizon

	Budget Level : Low		Budget Level : Medium		Budget Level : High	
	% Gap	% Improvement	% Gap	% Improvement	% Gap	% Improvement
FSSS	3.56	6.56	1.66	12.45	3.28	20.28

FSNS	3.87	12.66	1.67	16.32	6.48	35.36
VSSS	6.43	181.36	7.54	45.88	8.56	28.54
VSNS	9.78	266.54	9.33	47.36	9.34	33.28

Figures

Figure 1A: Impact of Prevalence Rate Reduction on Budget and QALYs gained



References

Anaya, H.D., Chan K., Karmarkar, U.S, Asch, S.M., Goetz, M.B. 2012. Budget Impact Analysis of HIV Testing in the VA Healthcare System, *Value in Health*, **15**, 1022-1028.

Freedberg, K. A., Scharfstein, J. A., Seage III, G. R., Losina, E., Weinstein, M. C., Craven, D. E., Paltiel, A. D. 1998. The cost-effectiveness of preventing AIDS-related opportunistic infections. *JAMA: the Journal of the American Medical Association*, ***279***(2), 130-136.

Gandhi, N. R., Skanderson, M., Gordon, K. S., Concato, J., Justice, A. C. 2007. Delayed presentation for human immunodeficiency virus (HIV) care among veterans: a problem of access or screening?*Medical Care*, **45**(11), 1105.

Kaplan, J.E., Benson, C., Holmes, K. K., Brooks, J. T., Pau, A., Masur H., 2009. Guidelines for prevention and treatment of opportunistic infections in HIV-infected adults and adolescents.*MMWR Recomm Rep*58, no. RR-4: 1-207.

Mauskopf, J., Kitahata, M., Kauf, T., Richter, A., Tolson, J. 2005. HIV antiretroviral treatment: early versus later. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, **39**(5), 562-569.

Paltiel, A. D., Weinstein, M. C., Kimmel, A. D., Seage III, G. R., Losina, E., Zhang, H., Freedberg, K. A., Walensky, R. P., 2005. Expanded screening for HIV in the United States—an analysis of cost-effectiveness.*New England Journal of Medicine*, **352**(6): 586-595.

Schackman, B. R., Gebo, K. A., Walensky, R. P., Losina, E., Muccio, T., Sax, P. E., Weinstein, M. C., Seage III, G. R., Moore, R. D., Freedberg, K. A., 2006. The lifetime cost of current human immunodeficiency virus care in the United States.*Medical Care,***44**(11) 990-997.