Health Care Manag Sci (2015) 18:363–375
DOI 10.1007/s10729-014-9304-0

# Applying reinforcement learning techniques to detect hepatocellular carcinoma under limited screening capacity

**Elliot Lee · Mariel S. Lavieri · Michael L. Volk · Yongcai Xu**

Received: 21 November 2013 / Accepted: 29 September 2014 / Published online: 12 October 2014


**Abstract** We investigate the problem faced by a healthcare system wishing to allocate its constrained screening resources across a population at risk for developing a disease. A patient's risk of developing the disease depends on his/her biomedical dynamics. However, knowledge of these dynamics must be learned by the system over time. Three classes of reinforcement learning policies are designed to address this problem of simultaneously gathering and utilizing information across multiple patients. We investigate a case study based upon the screening for Hepatocellular Carcinoma (HCC), and optimize each of the three classes of policies using the indifference zone method. A simulation is built to gauge the performance of these policies, and their performance is compared to current practice. We then demonstrate how the benefits of learning-based screening policies differ across various levels of resource scarcity and provide metrics of policy performance.

**Keywords** Screening · Reinforcement Learning · Liver Cancer · Simulation

E. Lee (✉) · M. S. Lavieri · Y. Xu
The University of Michigan, 1205 Beal Ave., Ann Arbor, MI 48109, USA
e-mail: elliotdl@umich.edu

M. S. Lavieri
e-mail: lavieri@umich.edu

Y. Xu
e-mail: yongcai@umich.edu

M. L. Volk
University of Michigan Health System, 1500 E. Medical Center Drive, Ann Arbor, MI 48109, USA
e-mail: mvolk@med.umich.edu

## 1 Introduction

Over six million Americans are estimated to be at risk for developing Hepatocellular Carcinoma (HCC) [44]. The incidence rate of HCC in the United States as of 2005 was 4.9 persons per 100,000, a rate which has tripled since 1975 [2].

Early detection is highly correlated to patient health outcomes; less than 10 % of the patients who are diagnosed with late-stage HCC survive beyond 5 years, whereas more than 50 % of the patients diagnosed with early-stage HCC are disease free after 5 years [38]. Therefore, the primary goal of HCC screening programs is to detect the development of the disease in the early stage.

We consider a screening program which has a limited screening capacity in each period. More precisely, the number of patients at risk for developing HCC outweighs the number of screenings available for administration in each period, and thus the problem of deciding which subset of the population to screen in each period arises.

This situation of a limited screening capacity could arise as a product of multiple scenarios. For instance, in highly overbooked screening clinics, the limited capacity results due to operational constraints of the clinics. Our approach addresses the challenge of finding a suitable screening program which accounts for their overbooked settings. Moreover, capacity constraints could arise as a decision maker is faced with the problem of improving population-wide health outcomes without using additional resources beyond the current expenditure, such as in third world countries. This approach also becomes more relevant in the face of soaring costs in American healthcare [7].

The current recommended screening protocol in the United States is to screen all at-risk patients every six months. The full definition of the at-risk population for

HCC is defined in [8]. In this paper, we restrict that definition to be those patients with chronic Hepatitis C with advanced fibrosis, which are two key risk factors for HCC.

The main disadvantage of fixed interval screening is that it does not take into account the information learned sequentially, which differentiates patients at various levels of risk of developing HCC. More intelligible behavior would entail allocating resources according to the risk learned.

Reinforcement learning algorithms are well-suited to handle these problems of sequential learning under constrained resources, as it will be demonstrated in this paper. It is our goal to provide insight into what types of behaviors are characteristic of efficient screening for Hepatocellular Carcinoma.

The paper proceeds as follows. We begin in Section 2 by providing the relevant background on HCC screening. In Section 3, we review relevant literature. In Section 4, we describe the details of the screening policies to be investigated. In Section 5, the details of the simulation through which the policies will be evaluated are given. We then use the indifference zone method in Section 6 to optimize the candidate screening policies through simulation. In Section 7, we provide the results of the best performing policies, as well as their comparison against current practice. We conclude by providing insight into what behaviors of efficacious screening can be extracted from this study in Section 8, along with avenues for future research.

## 2 HCC screening

Two things occur when a patient is screened for HCC. An ultrasound image of the patient's liver is taken and examined by a doctor. The doctor will order more accurate tests (such as CT or MRI) if any suspicious features in the ultrasound suggest the development of a tumor.

Secondly, the patient's blood is measured for the alphafeto-protein (AFP) level. The AFP is a biomarker which is weakly correlated with HCC. Given that this correlation is weak, the AFP is not explicitly utilized in treatment or screening decisions [11].

However, it has been shown that certain dynamics of the AFP, namely the standard deviation of a patient's AFP and the rate of rise of the patient's AFP, can significantly improve estimations of the patient's risk of developing HCC [24]. The following list contains risk factors identified in that study (where * indicates values measured upon enrollment into the surveillance program)

- Age*
- Black ethnicity*
- Blood platelet count*
- Ever having been a smoker*
- Alkaline phosphatase*
- Presence of esophageal varices*
- Standard deviation of all AFP readings while under surveillance
- Rate of AFP rise over time, determined by ordinary least squares estimate of all AFP readings while under surveillance

To illustrate the nature of the last two risk factors, Fig. 1 depicts sample AFP paths of patients in our dataset (which will be discussed thoroughly in Section 6.1) Patients who eventually develop cancer tend to be characterized by AFP measurements which are wildly fluctuating, while simultaneously trending upwards. On the other hand, patients who do not develop HCC tend to have more stable, predictable AFP measurements.

Intuitively, detection rates could be maximized by shifting resources towards high risk patients, away from low risk patients. However, re-allocating resources according to patients' risk is not a straightforward problem due to the fact

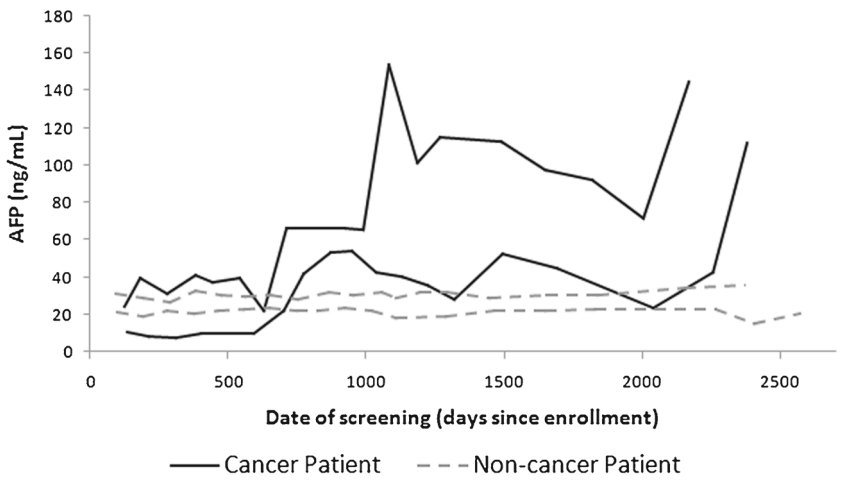


**Fig. 1** Sample AFP progressions for 2 patients who did and did not develop cancer

that, at any point in time, the decision maker holds imperfect knowledge of these dynamics. While baseline information about the patient provides some initial knowledge about the patient's risk of developing HCC, it is incomplete without understanding the patient's AFP dynamics. The challenge lies in the fact that knowledge of the AFP dynamics must be observed over multiple visits. Therefore, a sequential stage learning problem arises in the question of how to simultaneously gather and utilize this knowledge with limited resources.

## 3 Relevant literature

The mathematical analysis, design, and optimization of medical screening policies is expansive, and has seen significant proliferation in the past decade. Comprehensive surveys of this field include [1, 22, 39], and [34].

Research in the field varies in goals, such as the cost-effectiveness analysis of existing screening programs [16, 17, 27] the cost-effectiveness analysis of proposed hypothetical screening programs [12, 20, 23], and the optimization of new screening programs [19, 25, 26, 33, 35, 41]. Our work is of the latter type, as aforementioned recent medical discoveries have presented the opportunity for novel policies to be considered.

Previous research also varies in how the objective is defined. Authors seek to maximize quality-adjusted life years gained [4, 9, 15], minimize negative health outcomes of a patient [33, 41], or minimize costs to the system [16, 27, 31]. Others yet consider a combination of the above, usually providing a pareto-optimal set of policies, as seen in [18, 30, 36]. As mentioned in the previous section, the stage at which HCC is diagnosed is heavily correlated to the patient's likelihood of survival. We analyze policies with respect to two performance metrics: (1) the proportion of all cancers detected in the early stage, and (2) the proportion of screening resources spent on cancer patients.

Depending on the goal and problem setting, a large number of methodologies have been proposed, including Markov chains [17, 20, 23, 31, 35], simulation [10, 12, 16, 20], Markov decision processes [9, 27, 30], partially observable Markov decision processes [3, 5, 15, 46, 46], hidden Markov chains [30], and other stochastic models [25, 26, 36]. There are even less traditional methodologies used for analysis, such as game theory [45] and queueing theory [18]. For this work, simulation based optimization was chosen for its flexibility to analyze less traditional policies unable to be handled within the framework of the above methodologies.

Historically, researchers have often focused on simulating each patient's medical history by sampling from population parameters. This method is used by Loeve et. al. [28] in the simulation of colorectal cancer screening, and again by Urban et. al. [42] in the simulation of ovarian cancer screening. Our study differed primarily in choosing to use historical data to retroactively draw each patient's medical progression. Our approach has the advantage of having actual historical patients experience the proposed policy. It does not, however, have the robustness of a population-based simulation which can easily have a million unique, albeit fabricated, patients experience each policy.

Within the healthcare field, simulation based optimization has already seen much success. Zhang used a bisection search algorithm to find optimal capacity levels in long term care facilities [47]. Romero employed built-in optimization packages with the simulation software Arena to find optimal operational parameters of a skin cancer clinic [37]. Dhamodharan utilized Monte Carlo sampling methods, along with their developed simulation model, to optimize the implementation of immunization services in rural areas [13]. Simulation based optimization has not yet been used to optimize learning-based screening policies under constrained resources. We will rely on the indifference zone method to obtain optimal policies to address this question.

## 4 Problem setting

We consider a healthcare system with a panel size of $i = 1, ..., n$ patients at risk of developing HCC. All patients at time $t = 0$ are known to begin in a cancer-free state. In our problem, the size of the population outweighs the number of screenings available, and it is the task of the decision maker (DM) to decide which subset of the population to screen. At each time $t = 0, 1, 2, .., T$, the DM can choose a subset of $k < n$ patients to screen.

Each patient $i$'s risk of developing HCC can be measured by the following equation:

$$P(\text{HCC})_i = [1 + exp(-c_1 B_i - c_2 SD_i - c_3 RR_i)]^{-1} \quad (1)$$

- $P(\text{HCC})_i$ is the patient's lifetime cumulative probability of developing HCC,
- $B_i$ is a vector of all static risk cofactors measured upon enrollment into surveillance (age, ethnicity, smoker, alkaline phosphatase, blood platelets, and esophageal varices),
- $SD_i$ is the standard deviation amongst a patient's recorded AFP readings,
- $RR_i$ is the least squares estimate for the rate of AFP rise over time amongst a patient's recorded AFP readings,
- $c_1$ is a vector of the corresponding regression coefficients for all static risk factors, and
- $c_2$ and $c_3$ are regression coefficients for the AFP standard deviation and the rate of AFP rise over time.

Equation 1 calculates a patient's lifetime cumulative risk of developing HCC based upon several risk factors. For simplicity of notation, multiple risk factors which do not vary over time have been combined into a single quantity, $B_i$. The equation was determined through a nested case-control study in which risk factors for HCC development were assessed through conditional logistic regression [24].

The knowledge of the DM at time $t$ is captured by the state space variable:

$$(B_i, \hat{SD}_{i,t}, v_{i,t}, \hat{RR}_{i,t}, w_{i,t}) \quad (2)$$

$$\forall i = 1, .., n, \forall t = 0, 1, 2, ...T$$

Where the subscript $t$ has been added to emphasize that the DM only holds estimates of these quantities for each patient $i$ at each time $t$. $\hat{SD}_{i,t}$ is the sample standard deviation of all AFP observations for patient $i$ up to, and including, time $t$. $\hat{RR}_{i,t}$ is the rate of AFP rise over time for patient $i$, estimated by ordinary least squares of all AFP readings up to, and including, time $t$. Note that this quantity can be negative. $v_{i,t}$ is the variance of the standard deviation estimate $\hat{SD}_{i,t}$, and $w_{i,t}$ is the variance of the rate of rise estimate $\hat{RR}_{i,t}$, both calculated by standard statistical methods.

The DM can utilize Eq. 1 to obtain an estimate of patient $i$'s risk at time $t$ by using the equation:

$$\hat{P(\text{HCC})}_{i,t} = [1 + exp(-c_1 B_i - c_2 \hat{SD}_{i,t} - c_3 \hat{RR}_{i,t})]^{-1} \quad (3)$$

Note that we have simply adapted Eq. 1 by replacing unknown clinical quantities with sample estimates in order to approximate a patient's risk when said quantities are not perfectly known.

When the DM chooses to screen a patient $i$, two events occur in succession:

Firstly, the patient is revealed to be either (1) still cancer-free, (2) in early-stage cancer, (3) in late-stage cancer, or (4) dead, whether from cancer or other causes. If either outcome (2), (3), or (4) occur, the patient exits the system, and a new patient arrives in his/her place at time $t+1$, thus maintaining a constant panel size $n$.

Secondly, the patient's AFP level is measured. This additional reading is then used to re-estimate the AFP related state space variables for that patient: $\hat{SD}_{i,t+1}$, $v_{i,t+1}$, $\hat{RR}_{i,t+1}$, and $w_{i,t+1}$.

## 5 Reinforcement learning policies

### 5.1 Myopic behavior

An intuitive policy to investigate would be to act myopically upon current estimates $\hat{P(\text{HCC})}_{i,t}$. The algorithm proceeds as follows:

**Until** $t = T$

1. Rank the $n$ patients in non-descending order with respect to $x_{i,t} = (c_1 B_i + c_2 \hat{SD}_{i,t} + c_3 \hat{RR}_{i,t})$.
2. Screen the $k$ patients at the top of this list.
3. Advance to time $t + 1$.

Note that ranking patients according to $x_{i,t}$ is equivalent to ranking patients according to $\hat{P(\text{HCC})}_{i,t}$ because the function $(1 + exp(-x))^{-1}$ is monotonically increasing in $x$.

Naturally, this will utilize the DM's knowledge accumulated thus far to maximize the number of cancers detected in the current stage. The downside to this policy is that it fails to expand the DM's knowledge set for future decisions by exploring other patients. This behavior is often referred to as "pure exploitation" and usually performs suboptimally in various settings [40].

We consider three classes of reinforcement learning algorithms, each of which can be viewed as more intelligible modifications of this myopic behavior, with features to encourage exploration. We chose to study three distinct classes of algorithms to provide multiple perspectives on the value of learning within this problem setting, and thus increase the robustness of our conclusions.

### 5.2 $\epsilon$-greedy strategies

The first class of reinforcement learning algorithms that we consider are $\epsilon$-greedy strategies [43]. The algorithm (modified for our problem setting) proceeds as follows:

**Choose** $\epsilon \in [0, 1]$

**Until** $t = T$

1. Rank the $n$ patients in non-descending order with respect to $(c_1 B_i + c_2 \hat{SD}_{i,t} + c_3 \hat{RR}_{i,t})$.
2. Screen the $(1 - \epsilon) \cdot k$ patients at the top of this list.
3. Screen $\epsilon \cdot k$ patients from the remaining patients randomly.
4. Advance to time $t + 1$.

This strategy represents a slight modification of the myopic behavior as it reserves $\epsilon$ proportion of resources for exploration, and acts greedily with the remainder. It should be noted that myopic behavior is a special case of $\epsilon$-greedy strategies, corresponding to $\epsilon = 0$.

Conversely, the special case of $\epsilon = 1$ is often referred to as "pure exploration", where all choices are made randomly. Both pure exploration and pure exploitation provide benchmark performances for other reinforcement learning techniques to compare against.

### 5.3 Interval estimation strategies

The second class of reinforcement learning algorithms that we consider are interval estimation strategies [21]. These algorithms proceed as follows:

**Choose** $z \in [0, \infty)$
**Until** $t = T$

1. Rank the $n$ patients in non-descending order with respect to
   $c_1 B_i + c_2(\hat{SD}_{i,t} + z \cdot \sqrt{v_{i,t}}) + c_3(\hat{RR}_{i,t} + z \cdot \sqrt{w_{i,t}})$.
2. Screen the $k$ patients at the top of this list.
3. Advance to time $t + 1$.

Interval estimation is very similar in form to myopic behavior, but it encourages exploration by using a "distorted" risk score. Under very mild assumptions, patients whose risk of developing HCC is not currently known with high confidence are characterized by higher estimate variances, $v_{i,t}$ and $w_{i,t}$. Therefore by adding a multiplicative factor of the estimate variance to the risk score, the policy has artificially promoted patients with low knowledge up the list. The multiplicative factor, $z$, captures the incentive to explore as the relative importance of exploration. It should be noted that $z = 0$ corresponds to myopic behavior.

### 5.4 Boltzmann exploration strategies

The third class of reinforcement learning algorithms we consider are Boltzmann exploration strategies [29]. These algorithms proceed as follows:

**Choose** $\tau \in (0, \infty)$
**Until** $t = T$

1. Screen the $k$ patients, where patient $i$ is screened with probability $\frac{e^{x_{i,t}/\tau}}{\sum_{i'=1}^{n} e^{x_{i',t}/\tau}}$
   where $x_{i,t} = (c_1 B_i + c_2 \hat{SD}_{i,t} + c_3 \hat{RR}_{i,t})$, the original risk score.
2. Advance to time $t + 1$.

Boltzmann exploration gives all patients a positive probability of being screened, but with a probability which is weighted according to the DM's current risk estimates. The tuning parameter $\tau$ is known as the temperature. $\tau$ roughly captures the relative importance of exploration versus exploitation.

If $x_{i,t}$ is the original risk score, then the quantity $e^{x_{i,t}/\tau}$ can be thought of as a skewed risk score.

For example, when $\tau$ is very low, patients with only slightly differing original risk scores will have vastly differing skewed risk scores, due to the nature of the exponential function. Assessing patient risk according to the skewed risk score in this situation would be akin to artificially promoting exploitive behavior.

On the other hand, when $\tau$ is very high, the patients with vastly different original risk scores will have relatively similar skewed risk scores, again due to the nature of the exponential function. Therefore if we behave according to this skewed risk score, we have artificially promoted explorative behavior.

As $\tau \to \infty$, Boltzmann exploration approaches pure exploration.

## 6 Simulation

With three classes of candidate policies, a simulation was designed to serve as a testbed for empirical evaluation. The goal of this simulation was to receive proposed alternative screening policies as inputs, and then determine the number of cancers detected, as well as the resources used by each policy.

### 6.1 Description of the data

The Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) trial included 1050 patients followed for an average of 5.3 years. Surveillance of patients in this trial was performed in three ways: Firstly, the patients were screened every 3 months for the first 3.5 years, then every 6 months thereafter on a voluntary basis. At each screening visit, the level of alpha-fetoprotein (AFP) concentration in the blood was measured. Secondly, each patient underwent an ultrasound imaging approximately every 6-12 months. Thirdly, a liver biopsy was performed on all trial participants at 1.5 and 3.5 years into the trial. Table 1 displays the relevant characteristics of the patients in the HALT-C dataset used in our study.

**Table 1** Characteristics of patients with and without HCC

| Characteristic | HCC (N = 82) | NO HCC (N = 885) | *P* Value |
|---|---|---|---|
| Age at Baseline (years) | 53±7 | 50±7 | < 0.01 |
| Black (binary) | 24 % | 18 % | 0.10 |
| Platelets at Baseline $\times 1000/mm^3$ | 126±51 | 169±65 | < 0.01 |
| Ever Smoked (binary) | 41 % | 24 % | < 0.01 |
| Alkaline Phosphataste at Baseline (U/L) | 117±59 | 97±43 | < 0.01 |
| Esophageal Varices (binary) | 4 % | 34 % | 0.01 |
| Standard Deviation of AFP (ng/mL) | 51 ± 86 | 9 ± 19 | < 0.01 |
| Rate of AFP Rise (90*ng/mL) | 5±11 | 0.11±2.1 | < 0.01 |

For continuous variables, the mean ± standard deviation is shown, with *p*-values from a 2-sample t-test. For binary variables, the proportion is shown, with *p*-values from Fisher's exact test

Out of 1050 subjects, 83 were omitted from the current analysis for having $< 5$ AFP values available. Among the 967 subjects remaining, 82 developed HCC during the study period. During the screening period (time from enrollment to HCC diagnosis or end of follow-up), subjects had a median of 18 (range 5-23) AFP tests performed. It should be noted that the general American population has a lifetime risk for HCC of approximately 0.9 %. Our dataset demonstrated a much higher cumulative incidence due to the fact that the eligibility requirements of the HALT-C trial included a history of chronic hepatitis C with advanced fibrosis, a key risk factor for HCC. As HCC screening is not currently recommended for the general public, it is appropriate to study the performance of these policies on this at-risk subset population.

### 6.2 Model

In this simulation, we record three statistics:

1. $E$, the total number of early stage cancers detected during the planning horizon,
2. $L$, the total number of late stage cancers detected during the planning horizon, and
3. $X$, the number of screenings spent on patients who would eventually develop cancer

The discrete event logic is graphically depicted in Fig. 2. At all times, the simulation maintains two separate sets of patient data: (1) the simulation knowledge, and (2) the DM's knowledge. The latter is an incomplete subset of the former, which is further revealed through the DM's decisions of who to screen. The simulation begins at time $t = 0$ by using the Initial Panel Module to fill panel slots $i = 1, .., n$. Here, the simulation will randomly draw, with replacement, a patient history from the dataset to be this patient's simulated history. The result will be the creation of $C$, the set of patients who will develop cancer, and $N$, the set of patients who will not develop cancer in their lifetime. The DM's knowledge of these $n$ patients at this point is limited to the baseline score, $B$. Also at this time, the DM knows every patient to be cancer-free, as the dataset which we used was a clinical trial whose enrollment criteria included being cancer-free at the beginning of surveillance.

Next, the simulation runs the Policy Module, which receives the DM's knowledge of the current $n$ patients as an input, and the DM chooses a subset $K$ of patients to screen according to the current policy being evaluated. The value $X$ is then increased by the number of screenings which were spent on cancer patients, $|K \cap N|$.

For patients $i \notin K$ not chosen to be screened, the DM's knowledge of the patients will go unchanged until the next period. Patients chosen to be screened $i \in K$ enter the Imaging Module, which queries the simulation for the patient's current cancer state. If the patient never developed cancer during the course of HALT-C, the Imaging Module automatically outputs a cancer-free state. However, if the data indicates that this patient was detected to have a tumor of size $\bar{s}$ on date $\bar{t}$, we can estimate the tumor size $s$ on the simulated date $t$ according to the tumor doubling time $\delta$ [32] and the following doubling time equation:

$$s = 2^{\frac{t-\bar{t}}{\delta}} \cdot \bar{s} \tag{4}$$

The Imaging Module then assigns a cancer state according to the following logic:

$$\text{State} = \begin{cases} \text{Early} & \text{if } t \geq \bar{t} \text{ and } 1 \leq s \leq 5 \\ \text{Late} & \text{if } t \geq \bar{t} \text{ and } 5 < s \\ \text{Cancer-Free} & \text{if Otherwise} \end{cases} \tag{5}$$

It should be noted that the Imaging Module assumes that all tumors between 1 cm and 5 cm in size are detected with perfect accuracy. This assumption of perfect accuracy is supported by the fact that it is standard procedure to follow up any ultrasound which reveals suspicious features with

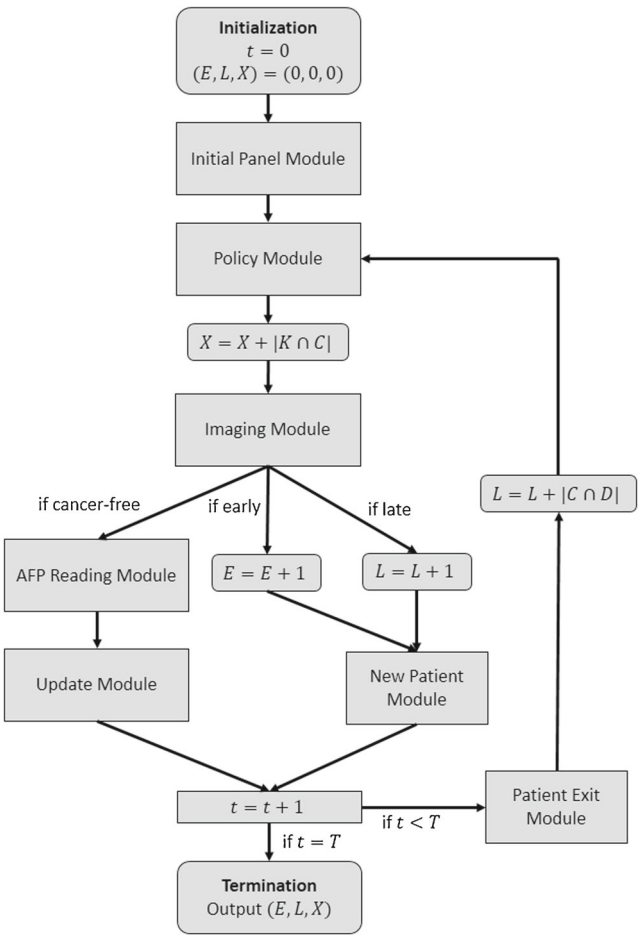


**Fig. 2** Discrete event simulation flow event logic

either a CT scan or MRI, both of which are highly accurate tests for tumor detection. This assumption could easily be relaxed in future work by incorporating the specificity and sensitivity of the respective tests.

The reason for assuming the tumor to be undetected on all dates $t < \bar{t}$ is to be as conservative as possible in our gauging in the performance of hypothetical policies. We could have utilized the same doubling time formula to determine the first date of tumor development, i.e the date at which the tumor was 1 cm in size. The tumor could theoretically be detected on any date after this first date of tumor development in the simulation. We instead adopted our more stringent definition in order to make it as difficult as possible for the investigated screening policies to outperform the real-world detection rates.

If the patient is cancer-free, then the simulation assigns a new AFP reading for this patient through the AFP Reading Module. On the simulated date, the dataset is queried for a linear interpolation between the two closest AFP readings in date. This simulated AFP reading is then added to the DM's knowledge by appropriately updating all state space variables to reflect this new reading in the Update Module.

If the patient is assigned an early-cancer state, the value $E$ is incremented by 1, and the patient is replaced according to the New Patient Module. The New Patient Module resets both the simulation knowledge of patient $i$ to a new patient drawn from the dataset, as well as the DM's knowledge of patient $i$. Furthermore, we assume the DM receives a single AFP reading for this patient.

Similarly, if a patient is assigned a late-cancer state, the value $L$ is incremented by 1, and the patient is replaced according to the same New Patient Module.

At the end of each period (with the exception of the final period), the Patient Exit Module is run to eliminate patients who depart the panel before the beginning of the next period. It does so by determining the departing subset $D \subset K$ whose time under surveillance in HALT-C has exceeded their duration in the simulation. These departures from surveillance include both outcomes of death or voluntary withdrawal from HALT-C. All patients who are eliminated are also replaced by a new patient before the beginning of the next period. In addition, all patients $i \in |D \cap C|$ who are eliminated through the Patient Exit Module also increment the penalty metric L by 1, as the policy has failed to identify a patient who had early stage cancer, and will now develop late-stage cancer outside of the simulated surveillance of that patient.

It should be noted that if the incoming replacement patient is drawn with equally likely probabilities, the population would become biased towards those patients with longer follow-up times. To maintain a patient panel which is probabilistically equivalent to that of HALT-C, the following probabilistic weights are used: if patients $j = 1, .., 967$ of HALT-C have follow-up times $p_j$, define $P = \sum_j p_j$. Patient $q$ is chosen for replacement with probability

$$\frac{\frac{P}{p_q}}{\sum_j \frac{P}{p_j}} \tag{6}$$

This process of deciding who to screen, and simulating the outcome of those screenings repeats until the end of the planning horizon $T$. Upon termination the three final values of $E$, $L$, and $X$ are reported to produce the following performance metrics:

1. The proportion of cancers detected in early stage $\frac{E}{E+L}$
2. The proportion of resources spent on patients who eventually develop cancer $\frac{X}{K \times T}$

### 6.3 Validation

Each iteration of the simulation begins by first randomly splitting the patients in the HALT-C data set into equally sized training and validation sets. The data in the training set is input into a conditional logistic regression to obtain coefficients $c_1$, $c_2$, and $c_3$ in Eq. 1. The patients in the validation set are used to populate the simulation. By this method, we avoid obtaining inflated estimates of policy performance, which would inevitably result by testing a policy on the same patients upon which the policy was built.

In testing the simulation outputs for agreement with real-world observations, the simulation predicted the number of early-cancers detected per year to be within 3 % of the results observed in the HALT-C dataset.

The model was built with high face validity by discussing the discrete-event logic alongside people involved in the screening process. Our co-author, a practicing clinician at the University of Michigan Hospital, helped to validate our model. We also interviewed the receptionists at the hospital responsible for booking screenings for patients, ultrasound technicians who perform the screenings, and nurses at the blood draw clinic responsible for measuring and reporting the AFP back to the doctor. These interviews were meant to strengthen our understanding of the real-world flow of events at every step of the screening process.

Lastly, we unilaterally deviated simulation parameters to extreme scenarios for "sanity checks", and checked for intuitive agreement with expected outputs.

## 7 Tuning parameter optimization

To find the optimal levels of tuning parameters within each class of reinforcement learning algorithms, we employed the indifference zone method of Dudewicz and Dalal [14]. This discrete optimization via simulation method considers

$m = 1, .., \ell$ discrete alternatives, where observations from population $m$ are normally distributed $N(\mu_m, \sigma^2{}_m)$.

The procedure begins by sampling each of the $\ell$ alternatives and equal number of times, via simulation. After the completion of this first stage, the sample variance of the simulation outcomes for each of the $\ell$ alternatives is calculated in order to determine the suitable number of additional samplings are needed for each alternative. After a second stage of sampling, a weighted average of the observations from the two stages is taken. The alternative $m$ with the highest weighted averaged is declared to be within $\delta$ of the true best with probability $P$.

The main advantage of this method over the ranking and selection method developed by Bechhofer [6] is that it does not require the assumption of $\sigma^2{}_1 = ... = \sigma^2{}_\ell := \sigma^2$, where $\sigma^2$ is known in advance. Initial analyses proved neither assumption to hold in this problem setting, thus encouraging the usage of the indifference zone method. Further details of this sampling procedure can be found in [14].

## 8 Results

### 8.1 Implementation parameters

Decision epochs were chosen to be at equally spaced 90 day intervals, under clinical recommendations that under no circumstance would it be necessary to again screen a patient less than 90 days after being screened. The planning horizon was chosen to be $T = 30$ years arbitrarily by the authors, although this analysis was first performed over 10 years, and the difference in results were insignificant. Finally, a panel size of $n = 500$ was chosen to mimic the approximate size of the screening program at the University of Michigan Hospital. The indifference zone method was run at a confidence of $P = 95\ \%$ with an indifference zone width of $\delta = 0.25$. These parameters are the result of initial analyses on the computation time required, and were chosen to maximize accuracy given the resources available.

The calculations were performed using MATLAB v2013a's Parallel Computing Toolbox, at the University of Michigan's Center for Advanced Computing, on 24 computing cores (intel i7, 4GB RAM). A single iteration of the simulation requires approximately 30 seconds of computing time. Our simulation was run for approximately 400 iterations per each of the 34 policies, per each of the 5 resource constraint levels.

### 8.2 Policy performance

The first analysis compared 5000 samples of current practice and pure exploration. Recall that pure exploration is equivalent to choosing a random subset of the population to screen in each period. Figure 3 is a histogram of the number of early cancers detected by these two policies. It is readily apparent that the two policies are highly similar in performance. This agrees with intuition, as both policies use no patient specific information, and treat all patients equally. Both policies can act as baselines to compare the performance of other policies against.

We optimized each of the three classes of learning-based policies at five settings of resource constraints with respect to the proportion of cancers detected in early stage. If we let $k$ be the number of screenings available to spend by the DM in each period, and $n$ be the size of the patient panel, then $\frac{k}{n}$ is the measure of how constrained the problem is. We analyzed this problem at $\frac{k}{n} = 0.10, 0.20, 0.30, 0.40,$ and 0.50 corresponding to five scenarios of varying resource scarcity.

$\epsilon$-greedy strategies were searched over the range $\epsilon = 0, 0.025, 0.050, 0.075, \ldots, 0.25$, interval estimation over $z = 1, 2, 3, ..., 10$, and Boltzmann exploration over $\tau = .250, 0.275, .300, .325, \ldots, .750$. The search ranges were chosen at the discretion of the authors, after some initial experiments to find suitable candidates, and observing significant drop-offs in performance beyond those bounds. We optimized each of the three classes of policies at each of the five resource constraint levels. Table 2 shows the results of the tuning parameter optimization.

Recall that for each of the three classes of policies studied, higher tuning parameters represent more emphasis upon exploration. From these results, we can see that as resources become less constrained, the optimal balance between exploration and exploitation shifts towards exploration. Conversely, as resources become more constrained, greater exploitation is encouraged. The tendency to explore less in highly resource constrained settings is intelligible, as

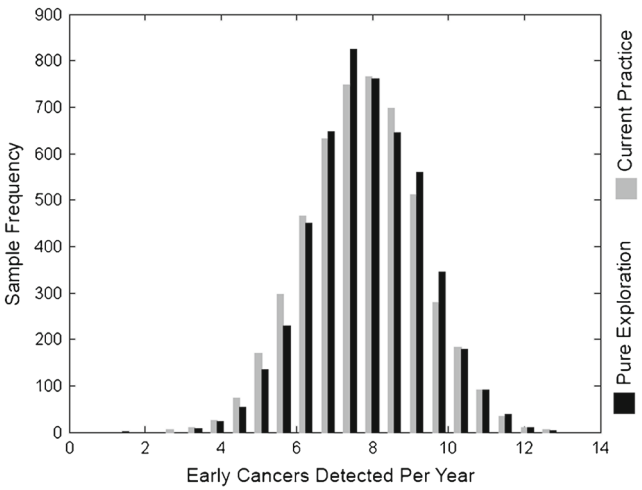


**Fig. 3** Comparison of performances of current practice and pure exploration

**Table 2** Optimal tuning parameters determined by the indifference zone method

| Resource constraint level $k/n$ | $\epsilon$-greedy $\epsilon$ | Interval estimation $z$ | Boltzmann exploration $\tau$ |
|---|---|---|---|
| 10 % | 0.025 | 1 | 0.250 |
| 20 % | 0.05 | 1 | 0.250 |
| 30 % | 0.10 | 1 | 0.300 |
| 40 % | 0.10 | 1 | 0.325 |
| 50 % | 0.25 | 5 | 0.400 |

the DM does not have enough resources to learn anything of significance. Therefore, in highly resource constrained settings, it would be prudent to depend more upon the baseline risk information received by the DM when the patient entered screening.

Next, we sought to compare the increase in performance of these optimized learning policies over current practice. Because the protocol of screening every patient every six months cannot be implemented in resource constrained settings, we created the equitable allocation policy to act as a policy equivalent in spirit to current practice. The equitable allocation policy uses whatever resources are available as fairly as possible, akin to current practice, by ensuring that all patients experience fixed interval screening at equal frequencies. Figure 4 displays our findings.

Our analysis determined Boltzmann exploration to be the policy which produced the most early stage cancer detections at every level of resource constraint. Furthermore, at every level of resource constraint, both myopic behavior and equitable allocation are dominated by any of the three reinforcement learning policies, thus demonstrating the importance of learning in our problem.

The most immediate benefit that can be drawn from these results is the increase in detection rates gained by switching to the best learning policy. Current practice screens 100 % of the population every 180 days, so it stands to reason that its performance is equivalent to equitably screening 50 % of the population every 90 days. The latter policy detects 58 % of patients in early stage cancer. The best performing learning policy reaches a 63 % detection rate at the same level of resource expenditure. This represents a 8.6 % increase in performance by switching from equitable allocation to the best performing learning policy.

Alternatively, we can analyze the cost-savings that can be achieved by switching to a learning based policy. The best performing learning policy only requires screening 41.75 % of the population every 90 days to achieve the same level of performance as current practice, which screens 50 % of the population every 90 days. This represents a 16.5 %

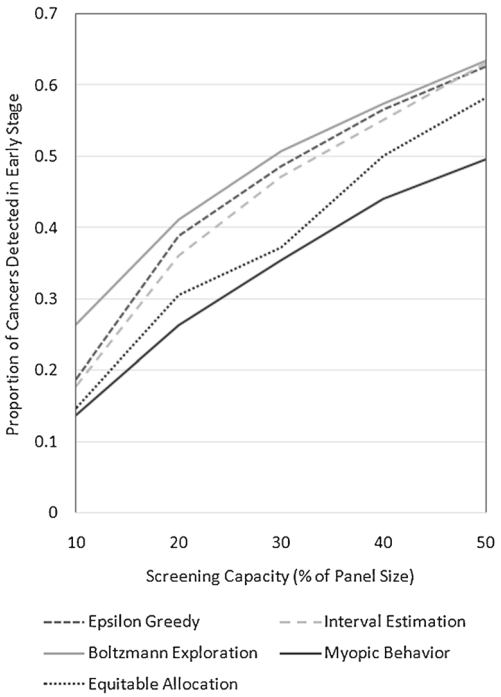


**Fig. 4** The performance of the optimal policies across various resource constraints

reduction in screening costs by switching from equitable allocation to the best performing learning policy, while maintaining the same level of performance.

As in many reinforcement learning applications, myopic behavior, or pure exploitation, is vastly suboptimal. This is due to the fact that myopic behavior can very easily become stuck in poor knowledge sets, and continue to incorrectly believe that certain subset of patients to be high risk. At 50 % resource constraint setting, the best learning policy has a 27 % increase in performance over myopic allocation of resources.

Another interesting feature of Fig. 4 is the relative gap between the learning policies and the equitable policy seems to decrease in size as more resources become available. This agrees with intuition because the more scarce a resource becomes, the more benefit there stands to be gained by acting efficiently. The value of learning policies can be made apparent by comparing the best performing policy at each resource level with equitable allocation in Fig. 5. Although the relative benefit of learning policies does generally decrease as more resources become available, it is still better than equitable allocation of resources.

We analyzed performance with respect to the percentage of screenings spent on cancer patients. This metric rewards a policy for screening the correct patients, even if those screenings did not immediately result in the detection of an early stage cancer. This metric is more concerned with the correct identification of high risk patients, and not

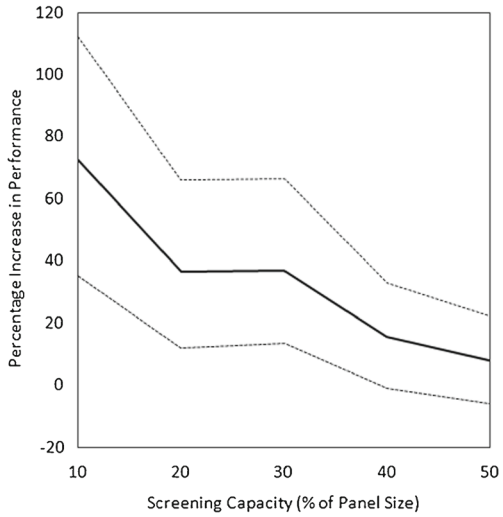


**Fig. 5** Increases in performance over equitable allocation of the best performing RL policy, with 90 % sampling percentiles, across various resource constraints

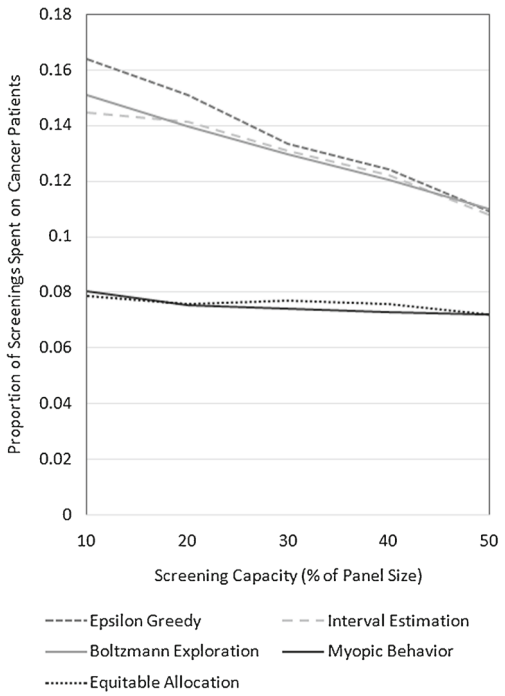


**Fig. 6** The proportion of screenings spent on cancer patients

necessarily the timing at which patients are screened. Although this metric is less useful in a clinical setting, it is actually more closely aligned to the original purpose of the reinforcement learning algorithms studied. The results are displayed in Fig. 6.

Equitable allocation spends roughly 8 % of its resources on cancer patients, across all resource levels. This is to be expected, as approximately 8 % of the HALT-C dataset develops cancer, and thus approximately 8 % of our simulated patient panel will eventually develop cancer. This metric more decisively demonstrates the advantage of learning-based policies over both equitable allocation, and myopic behavior. It also displays the decrease in relative benefit of learning-based policies as the capacity of the system increases.

Lastly, we investigated the effects of these policies from the patients' perspective. We sought to answer what a patient could expect to experience by participating in a screening regimen prescribed by our approach. We measured the 25th, 50th, and 75th percentile in days between subsequent screenings for each patient, then averaged these statistics across the population throughout the history of the simulation. These results are presented in Fig. 7.

From this figure we can glean what kinds of screening policies these reinforcement learning methods require a specific patient to undergo. There is a distinct gap between the screening frequencies experienced by patients who do and do not develop cancer. This holds for every type of policy, at every resource constraint level.

Under current practice, the screening capacity is 50 %, and both cancer and non cancer patients alike can expect to be screened once every 180 days. In the same setting, the average cancer patient being screened according to the optimal Boltzmann exploration policy can expect to have a median screening gap of 130 days. Similarly, the average patient who does not develop cancer will have a median screening gap of 296 days, an improved health outcome by means of avoiding unnecessary time, costs, and mental distress associated with screening visits. These same effects occur for all three policies.

We would advise a clinician looking to choose which of the three policies to adopt to choose based upon the findings in Fig. 7. If patient equity is a priority, epsilon greedy strategies seem to achieve the most similar screening gaps between cancer and non-cancer patients. If identification and treatment of cancer patients is a priority, then Boltzmann exploration policies are the best choice, as they give the most frequent screenings to cancer patients. On the other hand, if the avoidance of costs and hassle of non-cancer patients with unnecessary screenings is a priority, we would advise the clinician to adopt interval estimation policies, as they demonstrate a large advantage in infrequently screening non-cancer patients.

## 9 Discussion

In this work, we searched a large, but by no means exhaustive, class of reinforcement learning algorithms to evaluate the benefits that can be gained by reallocating the existing screening resources. We believe that this approach of

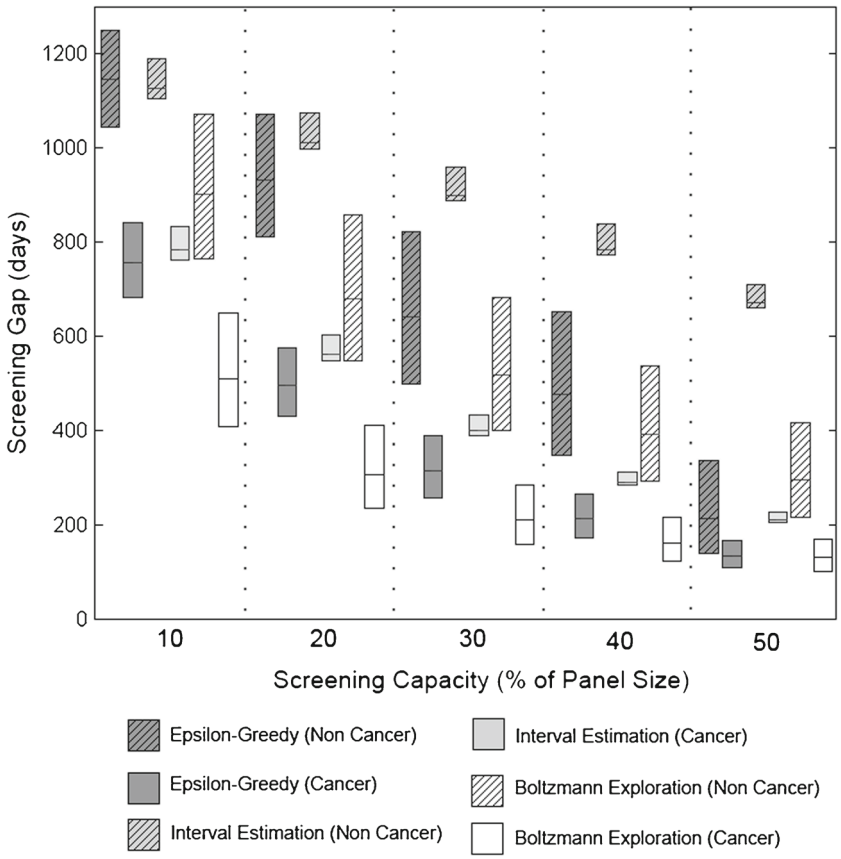


**Fig. 7** 25th, 50th, and 75th percentiles of screening gaps, averaged across cancer and non cancer patients separately

learning-based decision rules with a simulation built purely upon historical observations provides a highly accurate picture of the potential gains that can be made.

From this study, we induced several conclusions. The first is that current practice is roughly equal in performance to distributing resources randomly, thus creating the incentive to search for smarter behavior. Next, we saw that the optimal balance between explorative and exploitive behavior shifts towards the latter as resources become more scarce. We then estimated that switching from fixed-interval, equitable allocation of screening resources to a learning-based policy which utilizes sequentially gathered biological information will result in either 8.6 % increased performance or 16.5 % cost savings. We have also noted that these benefits of switching to a learning-based policy increase further as resources become more constrained. Lastly, we discussed the disadvantages of not utilizing learning based policies by showing myopic behavior to be vastly suboptimal.

We opted for the method of a historical simulation because of its potential for increased acceptance by the clinical community. This is due to two features of this method: we utilized a widely recognized clinical trials in the area of hepatocellular carcinoma, thereby creating results directly comparable to its well understood outcomes. Moreover, we avoided the usage of parametric assumptions on patient progression which are not recognized by the clinical community. These two features both establish strong rationale for a clinician to believe that our simulation accurately replicated their situation.

Nevertheless, we would like to note that our method of historical simulation is not without its shortcomings. The simulation may suffer from censoring which is inherent to the data from which we drew patient progressions. The simulation may also fail to accurately reflect a typical cohort of American patients with Hepatitis C and advanced fibrosis. These concerns are mitigated by the particularly robust nature of the HALT-C dataset. With patients remaining under surveillance for an average of 5.3 years, the impact of censoring is far less than that associated with a typical observational study. And with over 1,000 patients enrolled at 12 different hospitals from all regions of the United States, the HALT-C trial can be viably accepted as an accurate depiction of the American population living with chronic Hepatitis C and advanced fibrosis.

In our model, recall that a positive detection is simulated only if the tumor is determined to be both greater than 1 cm in diameter, as well as being beyond the date of detection in the original data. These detection rules were chosen to increase acceptance of our analysis by the medical

community. However, to evaluate the robustness of our results, a separate analysis re-sampled the best performing policies (which had been found based upon the original tumor detection assumptions) 5000 times each under an alternate set of assumptions where positive detections depended solely upon the size of the tumor.

We found that, although the performances of the policies were 15 % higher (averaged across all policies and scenarios), under these new detection rules our major findings continued to hold. That is, current practice could either gain 8.6 % increase in early stage detections, or a 16.4 % cost-savings, by switching to a reinforcement learning based policy for patients at risk for HCC. While this provides some evidence of the robustness of our results, it may be worthwhile in future work to derive the optimal reinforcement learning policies according to this alternate detection rule.

Further work may use stochastic simulation, where patient characteristics and their disease progressions are drawn parametrically. As an example, additive noise terms could be added to the AFP reading module to more realistically simulate the unpredictability of the AFP levels. Should the necessary parametrizations become established and available in the medical literature, this approach could potentially validate the results seen here, as well as provide further insight into the nature of efficient allocation of screenings.

These analyses could also be re-done with alternative objective functions to reflect other concerns of the screening clinic. Although we maximized the number of early-stage detections, it may be worthwhile consider rewards that are a function of the tumor size. Current staging definitions for HCC tumors use a threshold of 5 cm to distinguish between early stage and late stage tumors. While this may be a convenient definition for clinical classification, a patient's probability of survival may be better correlated with the tumor size at the time of his/her detection. This type of alternative objective function may identify a better policy more directly related to patient survival.

It might also be worthwhile to re-establish a measure of risk which turns all static risk factors (such as baseline age, smoking history, and baseline blood platelet count) into dynamic risk factors (current age, total years as a smoker, and current baseline blood platelet count). If this new measure of risk were to be established, we suspect it would only strengthen the decision making performed here. Our analysis could also benefit further from including the visual assessment of ultrasound images by doctors as an additional risk factor. Often a negative ultrasound will not contain enough features to warrant a diagnosis, yet it will still provide the doctor with some insight into the health of the liver organ. The main disadvantage of this usage of visual ultrasound images is that it is highly subjective between doctors, and it is difficult to quantify for a mathematical decision making framework.

Finally, other avenues for future work include complications that occur during real-life screening, such as false negative outcomes in the imaging process, panel size variability, penalties associated with screenings, and imperfect patient adherence.

We conclude with clinical recommendations derived from this work. Outside of direct policy adoption, a clinic can still utilize Eq. 1 in isolation to gauge their patient's estimated current level of risk. Furthermore, our results can be used for capacity planning purposes to gain a better understanding of the potential marginal benefits of increasing their current screening resources. Lastly, we would advise doctors to recognize the importance of balancing the exploration and exploitation of information when allocating resources.

**Acknowledgments** This material is based upon work supported by the National Science Foundation Graduate Student Research Fellowship under Grant No. DGE 1256260. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, in Ann Arbor, MI. The authors would also like to thank the editors, the three anonymous referees, as well as the guest reviewer for their thoughtful and constructive comments. Their suggestions proved to be very prudent, and were invaluable in the development of this paper.

## References

1. Alagoz O, Ayer T, Erenay FS (2011) Operations research models for cancer screening. Wiley Encyclopedia of Operations Research and Management Science
2. Altekruse SF, McGlynn KA, Reichman ME (2009) Hepatocellular carcinoma incidence, mortality, and survival trends in the united states from 1975 to 2005. J Clin Oncol 27(9):1485–1491
3. Ayer T, Alagoz O, Stout N (2009) A mathematical model to optimize breast cancer screening policy. In: Proceedings of the 31st annual meeting of the society for medical decision making abstract
4. Ayer T, Alagoz O, Stout NK (2012) OR Forum - A pomdp approach to personalize mammography screening decisions. Oper Res 60(5):1019–1034
5. Ayvaci MU, Alagoz O, Burnside ES (2012) The effect of budgetary restrictions on breast cancer diagnostic decisions. Manuf Serv Op Manag 14(4):600–617
6. Bechhofer RE (1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann Math Stat:16–39
7. Bodenheimer T, Chen E, Bennett HD (2009) Confronting the growing burden of chronic disease: can the US health care workforce do the job? Health Aff 28(1):64–74
8. Bruix J, Sherman M, Llovet J, Beaugrand M, Lencioni R, Burroughs A, Christensen E, Pagliaro L, Colombo M, Rodés J (2001) Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 easl conference. J Hepatol 35(3):421–430
9. Chhatwal J, Alagoz O, Burnside ES (2010) Optimal breast biopsy decision-making based on mammographic features and demographic factors. Oper Res 58(6):1577–1591

10. Clemen RT, Lacke CJ (2001) Analysis of colorectal cancer screening regimens. Health Care Manag Sci 4(4):257–267
11. Colli A, Fraquelli M, Casazza G, Massironi S, Colucci A, Conte D, Duca P (2006) Accuracy of ultrasonography, spiral ct, magnetic resonance, and alpha-fetoprotein in diagnosing hepatocellular carcinoma: a systematic review. Am J Gastroenterol 101(3):513–523
12. Davies R, Crabbe D, Roderick P, Goddard JR, Raftery J, Patel P (2002) A simulation to evaluate screening for helicobacter pylori infection in the prevention of peptic ulcers and gastric cancers. Health Care Manag Sci 5(4):249–258
13. Dhamodharan A, Proano R (2012) Determining the optimal vaccine vial size in developing countries: a Monte Carlo simulation approach. Health Care Manag Sci 15(3):188–196
14. Dudewicz EJ, Dalal SR (1975) Allocation of observations in ranking and selection with unequal variances. Sankhyā: Indian J Stat Series B:28–78
15. Erenay FS, Alagoz O, Said A (2014) Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. Manuf Serv Op Manag
16. Frazier AL, Colditz GA, Fuchs CS, Kuntz KM (2000) Cost-effectiveness of screening for colorectal cancer in the general population. JAMA: J Am Med Assoc 284(15):1954–1961
17. Goldie SJ, Kim JJ, Wright TC (2004) Cost-effectiveness of human papillomavirus dna testing for cervical cancer screening in women aged 30 years or more. Obstet Gynaecol 103(4):619–631
18. Güneś ED, Chick SE, Akśin OZ (2004) Breast cancer screening services: trade-offs in quality, capacity, outreach, and centralization. Health Care Manag Sci 7(4):291–303
19. Hanin L, Tsodikov A, Yakovlev AY (2001) Optimal schedules of cancer surveillance and tumor size at detection. Math Comput Model 33(12):1419–1430
20. Harper P, Jones S (2005) Mathematical models for the early detection and treatment of colorectal cancer. Health Care Manag Sci 8(2):101–109
21. Kaelbling LP (1993) Learning in embedded systems. MIT Press
22. Knudsen AB, McMahon PM, Gazelle GS (2007) Use of modeling to evaluate the cost-effectiveness of cancer screening programs. J Clin Oncol 25(2):203–208
23. Kulasingam SL, Benard S, Barnabas RV, Largeron N, Myers ER (2008) Cost effectiveness and resource. Cost Effectiveness and Resource Allocation 6:4
24. Lee E, Edward S, Singal AG, Lavieri MS, Volk M (2012) Improving screening for hepatocellular carcinoma by incorporating data on levels of $\alpha$-fetoprotein, over time. Clin Gastroenterol Hepatol 11(4):437–440
25. Lee S, Zelen M (2003) Modelling the early detection of breast cancer. Ann Oncol 14(8):1199–1202
26. Lee SJ, Zelen M (2008) Mortality modeling of early detection programs. Biometrics 64(2):386–395
27. Leshno M, Halpern Z, Arber N (2003) Cost-effectiveness of colorectal cancer screening in the average risk population. Health Care Manag Sci 6(3):165–174
28. Loeve F, Boer R, van Oortmarssen G, van Ballegooijen M, Habbema J (1999) The miscan-colon simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res 32(1):13–33
29. Luce RD (1959) Individual choice behavior a theoretical analysis. Wiley
30. Maillart LM, Ivy JS, Ransom S, Diehl K (2008) Assessing dynamic breast cancer screening policies. Oper Res 56(6):1411–1427
31. Myers ER, McCrory DC, Nanda K, Bastian L, Matchar DB (2000) Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. Am J Epidemiol 151(12):1158–1171
32. Okada S, Okazaki N, Nose H, Aoki K, Kawano N, Yamamoto J, Shimada K, Takayama T, Kosuge T, Yamasaki S (1993) Follow-up examination schedule of postoperative HCC patients based on tumor volume doubling time. Hepato-gastroenterology 40(4):311
33. Parmigiani G, Skates S, Zelen M (2002) Modeling and optimization in early detection programs with a single exam. Biometrics 58(1):30–36
34. Pierskalla WP, Brailer DJ (1994) Applications of operations research in health care delivery. Handbooks in OR & MS 6:4–7
35. Preston AJ, Smith W (2001) Disease screening designs: sensitivity and screening fruency. In: Proceedings of the annual meeting of the American statistical association, pp. 5–9
36. Rauner MS, Gutjahr WJ, Heidenberger K, Wagner J, Pasia J (2010) Dynamic policy modeling for chronic diseases: metaheuristic-based identification of pareto-optimal screening strategies. Oper Res 58(5):1269–1286
37. Romero H, Dellaert N, Geer S, Frunt M, Jansen-Vullers M, Krekels G (2013) Admission and capacity planning for the implementation of one-stop-shop in skin cancer treatment using simulation-based optimization. Health Care Manag Sci 16(1):75–86
38. Shaw FE (2010) Hepatocellular carcinoma - united states, 2001-2006. Centers for Disease Control and Prevention. Morb Mortal Wkly Rep 59(17):513–541
39. Stevenson C (1995) Statistical models for cancer screening. Stat Methods Med Res 4(1):18–32
40. Sutton RS, Barto AG (1998) Introduction to reinforcement learning. MIT Press
41. Tsodikov A, Szabo A, Wegelin J (2006) A population model of prostate cancer incidence. Stat Med 25(16):2846–2866
42. Urban N, Drescher C, Etzioni R, Colby C (1997) Use of a stochastic simulation model to identify an efficient protocol for ovarian cancer screening. Control Clin Trials 18(3):251–270
43. Watkins CJCH (1989) Learning from delayed rewards. Ph.D. thesis. University of Cambridge
44. Wilkins T, Malcolm JK, Raina D, Schade RR (2010) Hepatitis C: diagnosis and treatment. Am Fam Physician 81(11):1351–7
45. Yaesoubi R, Roberts SD (2008) How much is a health insurer willing to pay for colorectal cancer screening tests? In: Simulation Conference, 2008. WSC 2008. Winter, pp 1624–1631. IEEE
46. Zhang J, Denton BT, Balasubramanian H, Shah ND, Inman BA (2012) Optimization of prostate biopsy referral decisions. Manuf Serv Oper Manag 14(4):529–547
47. Zhang Y, Puterman ML (2013) Developing an adaptive policy for long-term care capacity planning. Health Care Manag Sci:1–9