# Generating Reward Functions using IRL Towards Individualized Cancer Screening

Panayiotis Petousis[1 [0000−0002−0696−608X]], Simon X. Han[1 [0000−0002−1001−4727]], William Hsu[1,2 [0000−0002−5168−070X]], and Alex A. T. Bui[1,2 [0000−0002−4702−1373]]

UCLA Bioengineering Department, Los Angeles, CA, 90095, USA
UCLA Department of Radiological Sciences, Los Angeles, CA, 90095, USA

**Abstract.** Cancer screening is a large, population-based intervention that would benefit from tools enabling individually-tailored decision making to decrease unintended consequences such as overdiagnosis. The heterogeneity of cancer screening participants advocates the need for more personalized approaches. Partially observable Markov decision processes (POMDPs) can be used to suggest optimal, individualized screening policies. However, determining an appropriate reward function can be challenging. Here, we propose the use of inverse reinforcement learning (IRL) to form rewards functions for lung and breast cancer screening POMDP models. Using data from the National Lung Screening Trial and our institution's breast screening registry, we developed two POMDP models with corresponding reward functions. Specifically, the maximum entropy (MaxEnt) IRL algorithm with an adaptive step size was used to learn rewards more efficiently; and combined with a multiplicative model to learn state-action pair rewards in the POMDP. The lung and breast cancer screening models were evaluated based on their ability to recommend appropriate screening decisions before the diagnosis of cancer. Results are comparable with experts' decisions. The lung POMDP demonstrated an improved performance in terms of recall and false positive rate in the second screening and post-screening stages. Precision $(0.02 - 0.05)$ was comparable to experts' $(0.02 - 0.06)$. The breast POMDP has excellent recall $(0.97 - 1.00)$, matching the physicians and a satisfactory false positive rate $(< 0.03)$. The reward functions learned with the MaxEnt IRL algorithm, when combined with POMDP models in lung and breast cancer screening, demonstrate performance comparable to experts.

**Keywords:** Cancer screening, maximum entropy inverse reinforcement learning, partially-observable Markov decision processes

## 1 Introduction

Annually, millions of people undergo screening for disease prevention and surveillance. From these tests, physicians aim to make decisions based on the patient's past results and most current observations, determining a subsequent action (e.g., further diagnostic testing, increased monitoring, following regular screening schedules, etc.) that optimizes early detection of health problems while balancing other (pragmatic) concerns (e.g., patient quality of life, resource utilization,

cost). Choosing the "best" next step and tailoring screening for each person is challenging: selecting an action of benefit in the immediate future may not be optimal over the long-term, given the particulars of an individual (i.e., a locally greedy approach vs. a global optimization).

Sequential decision making methods provide a potential solution. Such approaches can integrate and analyze multiple sources of patient data, while handling issues related to temporal credit assignment. In particular, partially observable Markov decision processes (POMDPs) have been applied to cancer screening (e.g., breast, colorectal, prostate [15]) to determine policies based on patients' risk factors and prior screening results. Markedly, POMDP models used in medicine typically use a reward function adopted from cost-effectiveness studies or are posed in terms of quality-adjusted life years (QALYs). While such functions are informative about general populations, they do not necessarily reflect how an experienced clinician would make a decision, especially given a specific individual's medical history and preferences. Indeed, little work has been done in designing reward functions that emulate experts' decision processes.

Here, we propose using the Maximum Entropy Inverse Reinforcement Learning (MaxEnt IRL) algorithm [18] to establish reward functions from retrospective screening data, learning how an expert physician may select a given action based on observed test results. We use an adaptive step size to expedite the convergence rate of MaxEnt IRL. Importantly, we present how to use the MaxEnt IRL learned rewards to generate state-action pair rewards that can be used in POMDPs. We demonstrate this work using two real-world clinical datasets for lung and breast cancer screening, mimicking how clinicians made decisions regarding patients. We evaluate the resultant POMDP policies using the MaxEnt IRL reward functions, comparing model performance to experts' actions. We conclude that the MaxEnt IRL algorithm is an efficient and accurate method in estimating sensible reward functions for cancer screening.

## 2 Background

Although Markov decision processes (MDPs) and POMDPs are used in a number of domains, their application in healthcare is limited and few strategies exist for estimating the associated reward functions that drive agent behavior in clinical settings. Classic examples include: Bennet et al. [4], who proposed a cost-effectiveness metric based on the cost required to obtain one unit of outcome change (CPUC); Hauskrecht et al. [8], who designed a reward model that combines economic cost and patient quality of life measures; and Tusch et al. [17], who predicated rewards on 30-day mortality risk for a surgical procedure. In contrast, we take advantage of growing amounts of longitudinal data, using recorded information and actions from electronic health records (EHRs) and other observational data sources to learn a POMDP reward function that imitates expert physicians' behavior for desired health outcomes. Specifically, IRL is proposed for this task.

Briefly, IRL addresses the problem of obtaining a reward function given an agent's optimal behavior over time towards a stated goal. A reward function for

the environment is unknown and is hence learned through empirical investigation of sensory inputs (i.e., observations) that progressively change the agent's selection of different actions. Two families of IRL algorithms exist: 1) linear programming (LP) methods [1,13]; and 2) probabilistic IRL algorithms [3,18]. While potentially more computationally complex, probabilistic IRL approaches have two advantages: they guarantee a unique solution for deterministic MDPs; and compared to LP methods, they can handle stochasticity in the data.Vroman et al. [3] developed a maximum likelihood IRL algorithm using clusters of experts' data trajectories to characterize different intentions. Applying the maximum likelihood IRL algorithm to each cluster subsequently derives a reward function representing the experts' behavior. Ziebart et al. [18] describe a probabilistic IRL algorithm that employs the principle of maximum entropy, dealing with noise and imperfect behavior as it normalizes globally over behaviors. In this approach, demonstrated for modeling routing preferences of vehicle drivers, behaviors with higher rewards are exponentially preferred by the algorithm when learning the reward function. Here, we build on and adapt this approach to obtain reward functions for cancer screening POMDPs.

## 3 Materials and Methods

### 3.1 NLST Dataset

The National Lung Screening Trial (NLST) is a multi-site randomized controlled trial that demonstrated a 20% mortality reduction in lung cancer screening using low-dose computed tomography (LDCT) relative to plain chest radiography [12]. For this work, we used data from the NLST's LDCT arm, comprising approximately 25,500 participants that underwent three annual screenings and follow-up post screening. We further filter this dataset to those subjects who had a reported pulmonary nodule based on imaging. Unfortunately, preprocessing of the NLST data is not straightforward, as longitudinal tracking of the nodules was not considered at the time of the study. Thus, to use imaging-related information, we made the assumption that an imaging finding in individuals with only one reported nodule and in the same anatomical location over time is the same nodule across the three screening points of the trial. This criterion further constrained our dataset to 5,694 LDCT subjects. From this subgroup, we learned a reward function, then trained and tested a POMDP. Note that for the reward function we made use of the recorded diagnostic follow-up variables (e.g., recommendation for other procedures) to inform actions.

### 3.2 Athena Dataset

The Athena Breast Health Network [7] is a University of California (UC)-wide initiative around breast cancer screening and treatment. The effort started in 2009 and includes women who underwent breast screening at five academic medical centers. The portion available at our institution (UCLA) consists of 49,244 patients, with follow-ups of up to 4.8 years; this subset represents 96,515 screening and diagnostic mammograms (MGs), and 2,713 diagnostic biopsies. MG

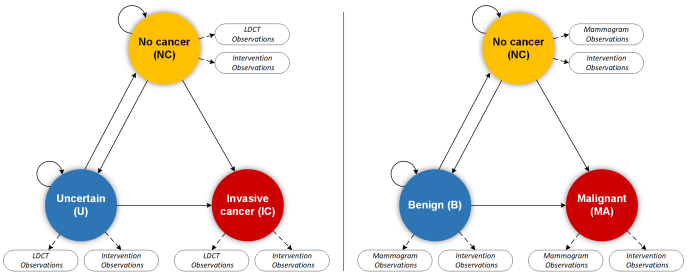



**Fig. 1. Left**. The lung POMDP; NC: no-cancer state; U: uncertain state; IC: invasive cancer state. LDCT and intervention observations can be observed in each state. **Right**. The breast POMDP; NC: non-cancer state; B: benign state; MA: malignant cancer state. MG and intervention observations can be observed in each state.

results are reported in Breast Imaging Reporting and Data System (BI-RADS) scores [6]. We selected patients with initial risk (Gail) scores, four consecutive screenings, and valid BI-RADS scores, which, along with biopsy results, per breast side (i.e., left, right). 2,095 patients with left breast MGs and 2,036 patients with right breast MGs (4,131 total cases) were used in this study.

### 3.3 Partially Observable Markov Decision Processes

An MDP is represented by a tuple of states, actions, rewards, action-dependent state transition dynamics (i.e., transition probabilities), and a discount factor. A POMDP is an extension to MDPs with two additional components: observations and state-dependent observation dynamics (i.e., observation probabilities). The state of the agent in POMDPs is partially observable. As such, its state is modeled as a probability distribution over the states, called the belief state, that gets updated over time based on the observations experienced by the agent.

We designed and evaluated two separate POMDPs for lung and breast cancer screening. Each model consists of three states and two actions. The observations of each POMDP are domain based: in the lung model, they represent findings obtained from LDCT imaging studies, including nodule size, consistency, location, and margins; in the breast model, they represent BI-RADS scores derived from MG interpretations. Given the nature of each dataset, both the lung and breast models have a horizon of three and four years, respectively, with 6-month and 1-year epochs. Each epoch represents time points for which we have information on the cancer status of patient (diagnosed with cancer or not). Transition and observation probabilities for each POMDP model are learned using the expectation maximization (EM) algorithm, for learning dynamic Bayesian networks, from each dataset. Both models were solved using the QMDP approximation solver [16].

**Lung cancer screening POMDP** Figure 1 (left) depicts the lung POMDP, illustrating the state space and allowed transitions between states, as well as the observations of each state. The state space consists of three states: the no-cancer (NC) state that represents any case with no suspicious abnormalities (i.e., no pulmonary nodules $>$ 4 mm). The uncertain (U) state that represents any

case with a noted finding (i.e., nodules 4mm or larger) but not yet a lung cancer. Lastly, the invasive-cancer (IC) state is any case with a confirmed lung cancer diagnosis through the use of additional diagnostic tests. The IC state is terminal such that any individual who enters it leaves the screening process for treatment. An LDCT action implies continuation of screening, whereas an intervention action refers to any diagnostic procedure (e.g., thoracotomy, biopsies, diagnostic CT, positron emissions tomography (PET) scan). Observations represent LDCT findings (nodule size, consistency, margins, and anatomic location) and the occurrence of an intervention. To generate initial belief states for each individual in our dataset we used the Tammemägi $\text{PLCO}_{\text{M2012}}$ model with demographic and clinical features at baseline to predict the risk of cancer. Demographic features used include age, education, race, and body mass index. Clinical features used were COPD, family history of lung cancer, personal history of cancer, smoking status, smoking intensity, and duration of smoking.

**Breast cancer screening POMDP** The breast POMDP model also consists of three states: the no-cancer (NC) state in which no abnormalities are seen, the benign (B) state in which benign breast disease diagnosis follows the MG, and the malignant (MA) cancer state in which the disease is confirmed through biopsy. MA is similarly a terminal state in which the patient leaves the screening process for treatment. Figure 1 (right) shows the breast cancer screening POMDP, transitions, observations (BI-RADS scores 1, 2, 3, 4A, 4B, 4C, 5), and actions. Though an intervention (biopsy in the breast cancer context) is possible after each MG, in practice biopsies are only performed after an MG of BI-RADS 4 or higher. For an initial belief, we used the patient's Gail score. The Gail score is an absolute risk estimate derived using age, age at menarche, age at first birth, the number of first-degree relatives with breast cancer, the number of previous breast biopsies, and race.

### 3.4 Maximum Entropy IRL

In IRL, the reward function, $r$, is assumed to be a linear combination of feature vectors $f_s$ and weights $\theta$ ($\theta^T$ is the transpose of $\theta$):

$$r(\tau;\theta) = \theta^T f_\tau = \sum_{s\in\tau} \theta^T f_s \tag{1}$$

A feature count, ($f_\tau$), is the sum of feature vectors of the states visited along a trajectory, where $f_s$ represents binary vectors indicating state values. Inputs to the MaxEnt IRL algorithm are an MDP and a set of trajectories ($D$) [2]. A path or a trajectory ($\tau$) represents the sequence of states ($s$) and ensuing actions followed by an agent in an MDP. For example, in the NLST dataset, a trajectory comprises three epochs (i.e., the three annual screening exams) with state-action pairs describing the lung cancer states and the actions taken (e.g., NC-LDCT, U-LDCT, and $\text{IC-I}_{\text{Biopsy}}$ ). The probability of a trajectory occurring in our set of trajectories is proportional to the exponential of the reward/cost of the trajectory [5]:

$$p(\tau;\theta) \propto \exp\left(r(\tau;\theta)\right) \tag{2}$$

As such, trajectories of equal reward are equally likely to be executed by the expert, whereas trajectories of less reward are less likely. The probability distribution over paths with maximum information entropy is parameterized over $\theta$. $Z(\theta)$ is the partition function, where $Z(\theta) = \sum_{\tau \in D} \exp r(\tau; \theta)$.

$$p(\tau; \theta) = \frac{1}{Z(\theta)} \exp\left(r(\tau; \theta)\right) \tag{3}$$

The log likelihood of the trajectories (loss function) is shown in Equation 4, $M$ is the number of trajectories:

$$L = \frac{1}{M} \sum_{\tau \in D} r(\tau; \theta) - \log \sum_{\tau \in D} \exp\left(r(\tau; \theta)\right) \tag{4}$$

This loss function is convex for a linear reward function and a deterministic MDP. To update $\theta$ we use a gradient descent function, where $\eta$ represents the learning rate:

$$\theta_{i+1} = \theta_i + \eta \nabla_\theta L \tag{5}$$

The gradient $\nabla_\theta L$ represents the difference of feature expectations and sum over state visitation frequencies multiplied with feature vectors:

$$\nabla_\theta L = \tilde{f} - \sum_{s_i} D_{s_i} f_{s_i} \tag{6}$$

A feature expectation, $(\tilde{f})$, is defined as the average of all feature counts across all trajectories. The frequency of state visitation, $D_{s_i}$, can be computed using a dynamic programming algorithm; see [2, 5] for more information regarding this algorithm. The pseudocode of the MaxEnt IRL algorithm can be found in [5].

### 3.5 Adaptive step size

To improve the convergence of the MaxEnt IRL algorithm, we introduce an adaptive learning rate approach for the update rule of the gradient descent. The idea behind making the step size adaptive is to calculate the inner product of $\nabla_\theta L$, the gradient, in the current step, i.e., $\nabla_\theta L_i$ with $\nabla_\theta L_{i-1}$, its value from the previous step. If the two are in the same direction then the step size can be increased, otherwise it is decreased. Following [10] we define the learning rate $\eta = \frac{\alpha}{(t+A)^\alpha}$, where $t$ is dependent on the gradient inner product (which becomes the dot product in higher dimensions); $\alpha$ and $A$ are constants. The role of $t$ is to regulate the learning rate:

$$t_{i+1} = \max(t_i + f(\langle -\nabla_\theta L_i, \nabla_\theta L_{i-1} \rangle), 0) \tag{7}$$

In this definition, $f(\cdot)$ represents the following sigmoidal function where $f(x) = f_{min} + \frac{f_{max} - f_{min}}{1 - \frac{f_{max}}{f_{min}} \exp -\frac{x}{\omega}}$. In the above expressions, $\alpha$, $A$, $f_{min}$, $f_{max}$, and $\omega$ are user-defined constants obtained from [10]. With $f_{min} < 0$, $f_{max} > 0$, and $\omega > 0$.

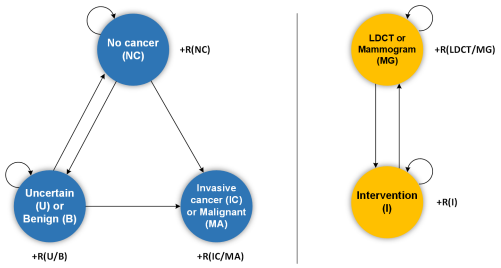



**Fig. 2. Left**. The state MDP; NC: non-cancer state; U/B: uncertain or benign state; I/MA: invasive or malignant cancer state, respectively for the lung and breast models. **Right**. The action MDP; LDCT/MG: state after a LDCT or MG; I: state after an intervention (e.g., biopsy); +R(·): rewards experienced by the agent in each state.

### 3.6 Computation of rewards

We assumed that given the outcome of a known cancer diagnosis for each individual over time, partial observability was no longer a problem while training, so learning the rewards of state-action pairs of an MDP instead of a POMDP was sufficient and computationally more efficient. However, the MaxEnt IRL algorithm computes the rewards of each state of an MDP, not state-action pair rewards ($r(s, a)$). To estimate rewards for each state-action pair combination, we designed two MDPs:

1. *A state MDP model.* The states of this MDP are the states depicted in Figure 2, for the lung and breast models. The transition matrix of the state MDP is the same transition matrix used in its respective POMDP model.
2. *An action MDP model.* In the action MDP, the states are defined by the previous action of the agent. These states model the options for screening (e.g., continue annual screening) and intervention (e.g., biopsy), in which the agent enters after performing each action. The action MDP transition model represents the probability of transitioning from the LDCT/MG state to the I state.

Figure 2 demonstrates the two MDPs. A combinatorial design decision inspired by [9] was used to learn state-action pair rewards. State-action pair rewards are computed using a multiplicative model shown in Equation 8:

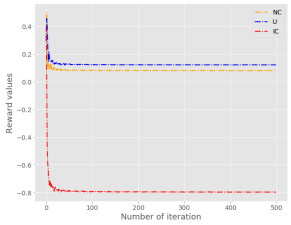
$$R(s, a) = R(s) \cdot R(a) \tag{8}$$

## 4 Evaluation and Results

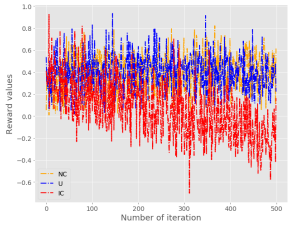
A stratified 5-fold cross validation study design was used to evaluate the POMDP models built from the NLST and the Athena datasets. The training set of each fold is used to learn the transition and observation matrices of the POMDPs, as well as the rewards using the MaxEnt IRL algorithm.

### 4.1 Comparison of MaxEnt IRL with & without adaptive step size

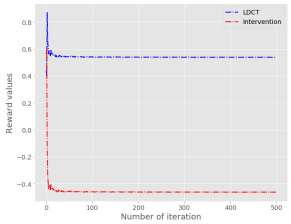
Table 1 shows the reward value of each state and action as well as different normalizations of these rewards computed using the MaxEnt IRL algorithm



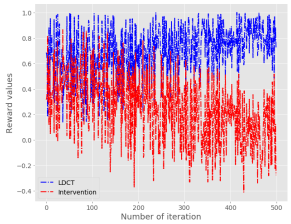

(a) Lung cancer states' rewards.




(b) Lung cancer states' rewards.




(c) Lung cancer actions' rewards.




(d) Lung cancer actions' rewards.

**Fig. 3.** State and action rewards computed using the MaxEnt IRL and normalized by range. **Left**: Using an adaptive step size. **Right**: Without using an adaptive step size. The adaptive step size MaxEnt IRL algorithm converges to a solution significantly faster than the MaxEnt IRL without an adaptive step size.

with an adaptive step size. We compare the MaxEnt IRL with and without the adaptive step size and assess the speed of convergence. Figure 3 depicts the computed rewards for states and actions for the lung POMDP over the number of iterations of gradient descent in the MaxEnt IRL algorithm, with and without an adaptive step size. A similar convergence trend is observed with the breast POMDP. As shown, the adaptive step size method converges to the correct solution more quickly than the standard MaxEnt IRL implementation. For the evaluation of the two models we use a reward function derived from rewards normalized in the [-1,1] range.

**Table 1.** The rewards for each state (R(NC), R(U/B), R(IC/MA)) and action (R(LDCT/MG), R(I)) computed using the MaxEnt IRL algorithm, for one of the folds of the 5-fold cross validation, with an adaptive step size.

| Normalization | R(NC) | R(U/B) | R(IC) | R(LDCT/M) | R(I) |
|---|---|---|---|---|---|
| **Lung cancer** | | | | | |
| **None** | 83.530 | 127.410 | -835.730 | 497.610 | -427.530 |
| **By range** | 0.080 | 0.120 | -0.800 | 0.540 | -0.460 |
| **[0,1]** | 0.950 | 1.000 | 0.000 | 1.000 | 0.000 |
| **[-1,1]** | 0.910 | 1.000 | -1.000 | 1.000 | -1.000 |
| **Breast cancer** | | | | | |
| **None** | -37.930 | 103.950 | -571.420 | -0.840 | -1179.820 |
| **By range** | -0.050 | 0.150 | -0.800 | -0.001 | -0.999 |
| **[0,1]** | 0.790 | 1.000 | 0.000 | 1.000 | 0.000 |
| **[-1,1]** | 0.580 | 1.000 | -1.000 | 1.000 | -1.000 |

### 4.2 Lung and breast POMDP results

We used the longitudinal observations from the NLST and Athena datasets as input to POMDPs such that each sequential observation updates the belief state of the agent. The belief state of the POMDP, at each epoch, is then used to select the next (optimal) action, with the objective of early detection of cancer. The POMDP models can suggest to continue screening (i.e., MG, LDCT) or to perform an intervention (i.e., biopsy or diagnostic imaging). If an intervention is performed, the individual is removed from further consideration. Evaluation of the POMDP is posed as a binary problem: if the POMDP suggests continued screening (LDCT/MG) then the patient is classified as a *negative* cancer; if it suggests an intervention, then the patient is classified as a *positive* cancer. Based on this definition, if the model suggests a LDCT/MG and the patient did not have a confirmed diagnosis of cancer in a given epoch, it is considered a true negative (TN); if the patient had a confirmed diagnosis of cancer then it is a false negative (FN). Conversely, if the model suggests an intervention and the patient did not have cancer in a given epoch, then it is considered a false positive (FP); if the patient had a diagnosis of cancer then it is considered a true positive (TP). Performance metrics were estimated for each epoch of the screening process. Any subject diagnosed with cancer is removed from the subsequent epoch. The POMDP models are compared against the equivalent physician decisions (recommendations) at each epoch, applying a similar framework for TN/FN/FP/TP to the experts, given the known cancer outcomes from each dataset (e.g., if the physicians suggested an LDCT/MG and the patient did not have a confirmed diagnosis of cancer, it is considered a true negative, etc.). Table 2 shows the performance of the lung and breast POMDPs and the corresponding performance of physicians on the same dataset. Notably, both POMDP models show performance comparable to experts. The lung cancer screening model has worse performance in terms of recall in the first and third screening epochs, but an improved performance in terms of recall and false positive rate in the second screening and post-screening. The breast cancer screening model demonstrates excellent recall (as do the expert physicians) but slightly worse false positive rate. The Cohen's kappa coefficient of agreement was used to assess the concordance between the POMDP models and physicians. The kappa score of the lung POMDP and physicians decreases over time due to the large number of false positives. A large portion of different cases are classified as false positives between the lung POMDP and physicians. The breast POMDP has a high kappa score demonstrating strong agreement with physicians in terms of false positives and true positives. For both lung and breast models, the variance of kappa per screening is less than 0.03.

## 5 Discussion

POMDPs, through the use of beliefs and a hidden state space, can overcome some of the limitations seen in other sequential decision making models used in cancer screening. For instance, we modeled a hidden cancer state space in three parts

**Table 2. Left**: The lung and breast POMDPs performance per epoch. **Right**: The physicians performance at each epoch. Metrics used for this evaluation are the true positive rate ($TP$), false negative rate ($FN$), false positive rate ($FP$) true negative rate ($TN$), precision ($P$), and recall ($R$). NCs: no-cancer cases. Cs: cancer cases. Kappa: Cohen's kappa score (coefficient of agreement), variance of kappa for all scores: $< 0.03$.

| | POMDP | | | | | | Physicians | | | | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Lung cancer** | | | | | | | | | | | | | |
| | **TN rate** | **FP rate** | **FN rate** | **TP rate** | **Precision** | **Recall** | **TN rate** | **FP rate** | **FN rate** | **TP rate** | **Precision** | **Recall** | |
| **Training** | **NCs: 4192 , Cs: Scr1, 2, 3 = 130, 68, 86 ; Pst-Scr = 78** | | | | | | | | | | | | |
| **Scr 1** | 0.48 | 0.52 | 0.02 | 0.98 | 0.05 | 0.98 | 0.48 | 0.52 | 0.00 | 1.00 | 0.06 | 1.00 | 0.42 |
| **Scr 2** | 0.34 | 0.66 | 0.02 | 0.98 | 0.02 | 0.98 | 0.34 | 0.67 | 0.05 | 0.95 | 0.02 | 0.95 | 0.29 |
| **Scr 3** | 0.24 | 0.76 | 0.01 | 0.99 | 0.03 | 0.99 | 0.21 | 0.79 | 0.00 | 1.00 | 0.03 | 1.00 | 0.05 |
| **Pst-Scr** | 0.25 | 0.75 | 0.07 | 0.93 | 0.02 | 0.93 | 0.22 | 0.78 | 0.14 | 0.86 | 0.02 | 0.86 | 0.05 |
| **Testing** | **NCs: 1048 , Cs: Scr1, 2, 3 = 32, 17, 21 ; Pst-Scr = 20** | | | | | | | | | | | | |
| **Scr 1** | 0.48 | 0.52 | 0.04 | 0.96 | 0.05 | 0.96 | 0.48 | 0.52 | 0.00 | 1.00 | 0.06 | 1.00 | 0.42 |
| **Scr 2** | 0.35 | 0.65 | 0.02 | 0.98 | 0.02 | 0.98 | 0.33 | 0.67 | 0.05 | 0.95 | 0.02 | 0.95 | 0.30 |
| **Scr 3** | 0.25 | 0.75 | 0.05 | 0.95 | 0.03 | 0.97 | 0.21 | 0.79 | 0.00 | 1.00 | 0.03 | 1.00 | 0.07 |
| **Pst-Scr** | 0.25 | 0.75 | 0.07 | 0.93 | 0.02 | 0.93 | 0.22 | 0.78 | 0.14 | 0.86 | 0.02 | 0.86 | 0.06 |
| **Breast cancer** | | | | | | | | | | | | | |
| **Training** | **NCs: 2908 , Cs: Scr1, 2, 3, 4 = 370, 68, 27, 5** | | | | | | | | | | | | |
| **Scr 1** | 0.99 | 0.01 | 0.01 | 0.99 | 0.96 | 0.99 | 0.99 | 0.01 | 0.01 | 0.99 | 0.95 | 0.99 | 1.00 |
| **Scr 2** | 0.99 | 0.01 | 0.01 | 0.99 | 0.70 | 0.99 | 0.99 | 0.01 | 0.01 | 0.99 | 0.73 | 0.99 | 0.97 |
| **Scr 3** | 0.98 | 0.02 | 0.03 | 0.97 | 0.40 | 0.97 | 0.99 | 0.01 | 0.03 | 0.97 | 0.43 | 0.97 | 0.95 |
| **Scr 4** | 0.98 | 0.02 | 0.00 | 1.00 | 0.09 | 1.00 | 0.98 | 0.02 | 0.00 | 1.00 | 0.10 | 1.00 | 0.92 |
| **Testing** | **NCs: 728 , Cs: Scr1, 2, 3, 4 = 93, 17, 7, 1** | | | | | | | | | | | | |
| **Scr 1** | 0.99 | 0.01 | 0.01 | 0.99 | 0.96 | 0.99 | 0.99 | 0.01 | 0.01 | 0.99 | 0.99 | 0.99 | 1.00 |
| **Scr 2** | 0.99 | 0.01 | 0.01 | 0.99 | 0.70 | 0.99 | 0.99 | 0.01 | 0.01 | 0.99 | 0.74 | 0.99 | 0.97 |
| **Scr 3** | 0.99 | 0.01 | 0.03 | 0.97 | 0.40 | 0.97 | 0.99 | 0.01 | 0.03 | 0.97 | 0.44 | 0.97 | 0.95 |
| **Scr 4** | 0.98 | 0.02 | 0.00 | 1.00 | 0.09 | 1.00 | 0.98 | 0.02 | 0.00 | 1.00 | 0.10 | 1.00 | 0.91 |

[14]: no-cancer, benign/indeterminate, and malignant/invasive cancer. Modeling the cancer state space with an additional state rather than a binary state space allows the distinction of lower risk individuals (i.e., no abnormalities) – who constitute a large portion of screening cases and thus result in highly imbalanced datasets – over medium (i.e., benign growth) and high risk individuals (i.e., malignant abnormality).

Driven by the need to define the reward function in these screening POMDPs, we explored the use of the MaxEnt IRL algorithm towards generation of state-action reward pairs. As noted earlier, cost and utility estimation are frequently adopted as reward functions in healthcare models. However, cost has certain limitations as it does not generalize to the whole population equally, and does not reflect the importance of quality outcomes. Additionally, QALY data are scarce, and arguably expensive to collect [11]. In contrast, a reward function learned using the MaxEnt IRL algorithm aims to maximize the objective of state-action trajectories. We introduced a multiplicative model for representing state-action pairs as products of state rewards and action rewards. The multiplicative model has the advantage to clearly demonstrate the difference in utility between rewards of different actions, which is what drives decision recommendation. Rewards are thus learned based on the state-visitation frequency of each trajectory. In this context, states with fewer visitations across each trajectory earn the lowest reward (e.g., invasive or malignant cancer state), which is why only cancer and non-cancer cases with a complete trajectory are used to learn rewards in our framework. Modeling the expert's decisions with the MaxEnt IRL algorithm resulted in reward functions for the POMDP models with performance comparable to experts. We noticed that when using aggressive reward functions (i.e., identifying all cancer cases), the true positive rate exceeded physicians' true

positive rate but at the expense of a higher false positive rate, which in clinical practice can translate into higher costs and unnecessary psychological burden on the patient. Including more observational variables, derived from medical images, in the screening process can overcome this trade-off between true positive and false positive rate. The overall true positive rate and false positive rate using our learned reward functions in the POMDPs is comparable to experts. Nonetheless, in some cases the experts had false negative cases, which is also captured by our approach. When compared with other machine learning algorithms at the baseline of the lung and breast paradigms the POMDP models demonstrate improved performance.

The first limitation of using MaxEnt IRL in this study is the fact that more than one combination of rewards can define the same problem to overcome this problem. To overcome this, a policy iteration algorithm can be used rather than value iteration as the policy space is finite in comparison to the rewards space (hence the policy iteration algorithm is guaranteed to optimally converge). A second limitation is the assumption that reward functions are only based on state visitation frequencies. To assess the quality of these reward functions a comparison of suggested recommendations with patient satisfaction could be used. Other limitations are around assumptions about the nature of our datasets. While lung and breast cancer screening tests occurred roughly at one year intervals, we assumed that screening occurs annually (i.e., at fixed frequency). Moreover, data imbalance is a function of time, as at each screening point the number of cancer and non-cancer cases changes (i.e., at the outset of a screening period, more cancers are found at the beginning of a dataset). We did not account for this dynamic nature of the dataset during training. Given the small number of cancer cases across each screening point of both datasets, we utilized a stratified 5-fold cross-validation to obtain an unbiased estimate of model performance. To simplify modeling, our lung POMDP model considered only cases reporting a single pulmonary nodule over the course of the trial; this represents only a subset of the screened individuals, as many subjects have more than one such finding. Lastly, for the Athena dataset, in breast cancer screening, patients with BI-RADS 1, 2, or 3 rarely undergo biopsy, thus the true FN rate is likely underestimated. Future work involves the exploration of MaxEnt IRL in transfer learning between other datasets and domains, by reusing learned weights.

## Acknowledgements

The authors thank the National Cancer Institute (NCI) for access to the National Lung Screening Trial data and Dr. Arash Naeim for access to the Athena Breast Health Network data collected at our institution. This material is based upon work supported by the National Science Foundation under Grant No. 1722516 and the Department of Radiological Sciences under the Data-Driven Diagnostic Decision Support (D4S) initiative.

## References

1. Abbeel, P., Ng, A.Y.: Apprenticeship learning via inverse reinforcement learning. In: Twenty-first international conference on Machine learning - ICML '04. p. 1

(2004). https://doi.org/10.1145/1015330.1015430
2. Alger, M.: Deep Inverse Reinforcement Learning. Tech. rep. (2016), https://matthewja.com/pdfs/irl.pdf
3. Babeş-Vroman, M., Marivate, V., Subramanian, K., Littman, M.: Apprenticeship learning about multiple intentions. Proceedings of the 28th International Conference on Machine Learning, ICML 2011 pp. 897–904 (2011)
4. Bennett, C.C., Hauser, K.: Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. Artificial Intelligence in Medicine **57**(1), 919 (2013). https://doi.org/10.1016/j.artmed.2012.12.003
5. Chelsea Finn: Deep RL Bootcamp Lecture 10B Inverse Reinforcement Learning - YouTube (2017), https://www.youtube.com/watch?v=d9DlQSJQAoI&t=1012s
6. D'Orsi, C.J.: ACR BI-RADS atlas: breast imaging reporting and data system. American College of Radiology (2013)
7. Elson, S., Hiatt, R., Anton, C.: The Athena breast health network: Developing a rapid learning system in breast cancer prevention, screening, treatment, and care, vol. 140. Springer US (7 2013). https://doi.org/10.1007/s10549-013-2612-0
8. Hauskrecht, M., Fraser, H.: Planning treatment of ischemic heart disease with partially observable Markov decision processes. Artificial Intelligence in Medicine **18**(3), 221–244 (2000). https://doi.org/10.1016/S0933-3657(99)00042-1
9. Hauskrecht, M., Milos, H.: Dynamic decision making in stochastic partially observable medical domains: Ischemic heart disease example. In: Lecture Notes in Computer Science, pp. 296–299. springerlink. https://doi.org/10.1007/bfb0029462
10. Klein, S., Pluim, J.P., Staring, M., Viergever, M.A.: Adaptive stochastic gradient descent optimisation for image registration. International Journal of Computer Vision **81**(3), 227–239 (2009). https://doi.org/10.1007/s11263-008-0168-y
11. Maillart, L.M., Ivy, J.S., Ransom, S., Diehl, K.: Assessing Dynamic Breast Cancer Screening Policies. Operations Research **56**(6), 1411–1427 (2008). https://doi.org/10.1287/opre.1080.0614
12. National Lung Screening Trial Research Team, Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., Clapp, J.D., Fagerstrom, R.M., Gareen, I.F., Gatsonis, C., Marcus, P.M., Sicks, J.D.: Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine **365**(5), 395–409 (8 2011). https://doi.org/10.1056/NEJMoa1102873
13. Ng, A.Y., Russell, S.: Algorithms for inverse reinforcement learning. Proceedings of the Seventeenth International Conference on Machine Learning pp. 663–670 (2000). https://doi.org/10.2460/ajvr.67.2.323
14. Petousis, P., Han, S.X., Aberle, D., Bui, A.A.: Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network. Artificial Intelligence in Medicine **72**, 42–55 (2016). https://doi.org/10.1016/j.artmed.2016.07.001
15. Schaefer, A.J., Bailey, M.D., Shechter, S.M., Roberts, M.S.: Modeling medical treatment using Markov decision processes. Operations Research and Health Care pp. 597–616 (2005). https://doi.org/10.1007/1-4020-8066-2_23
16. Thrun, S., Burgard, W., Fox, D.: Probabilistic robotics (2006). https://doi.org/10.1145/504729.504754
17. Tusch, G.: Optimal sequential decisions in liver transplantation based on a POMDP model. In: ECAI. pp. 186–190 (2000)
18. Ziebart, B.D., Maas, A., Bagnell, J.A., Dey, A.K.: Maximum Entropy Inverse Reinforcement Learning. In: AAAI Conference on Artificial Intelligence. pp. 1433–1438 (2008)